TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Proteomik und Bioanalytik

isobarQuant: A software tool for the processing, analysis and quantification of Mass Spectrometry Data

Toby Mathieson

Vollständiger Abdruck der von der Fakultät TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Dmitrij Frishman
Prüfer der Dissertation:	1. Prof. Dr. Bernhard Küster 2. Priv.-Doz. Dr. Christian Fufezan

Die Dissertation wurde am 27.04.2020 bei der Technischen Universität München eingereicht und durch die Fakultät TUM School of Life Sciences am 30.09.2020 angenommen.

Abstract

The volume of mass spectrometry-derived data is currently growing rapidly thanks to new acquisition methods and ever faster mass spectrometers. A large number of software tools to process this deluge of data exist, but many are limited to a small part of the complete analysis or are specialized in only answering certain experimental questions. Only a few programs are able to handle all aspects of a detailed analysis and at the same time allow access to all data in order to address further and specialized questions.

In recent years the field of mass spectrometry has also experienced a shift away from simply detecting the absence or presence of a protein in a cell to performing comparisons in the abundance of proteins between the normal state of a biological system and the effect of disease or a given substance; or to monitoring the changes in protein abundance over time. This proteome quantification requires methods that are very accurate and precise.

The aim of this PhD project was to craft a program that would be able to support this change and perform the necessary accurate and precise quantification. It should be easy to use and should bring together raw data and its interpretation into one place. This tool would build on and extend existing computational methods to process mass spectrometry data and make them available to the wider mass spectrometry community, allowing researchers to answer questions that had not yet been possible.

Following a general introduction to the topic, the second part of this thesis outlines the development of this tool, isobarQuant. It highlights its structure, layout and fundamental aspects of how it works; showcasing the processing and extraction of acquired raw data up to the point of the Mascot search and then further describes the internalization of the search output adjacent to the acquired data. Details about how it performs isobaric tag- based quantification are given and explanation is given as to how it can circumvent issues associated with that type of quantification.

The third part of this thesis describes how the isobarQuant tool is used and its capabilities extended to enable accurate and precise quantification in five different, non-dividing cell types using a peptide-ion based (SILAC) approach. The methods developed to enable this are presented: the use of an exact isotope model for the peptide quantification and the optimization of filtering based on a novel metric (prior ion ratio), alongside others, to achieve maximum coverage with highest accuracy and precision. This resulted in a publicly-available catalog of more than 9,600 protein half-lives in four human and one murine primary cell type and permitted the investigation into the longevity of protein complexes and how this varies across the different cell types.

The fourth part of this thesis makes use of the main isobarQuant output to investigate and profile the changes in peptide fragmentation brought about by the addition

of the TMT isobaric quantification tag. The change in fragmentation patterns is discussed in the context of how it can affect the score output from the Mascot search engine. Additionally, a new way of revealing potentially co-eluting peptides is suggested as well as an improvement to the existing H-score algorithm

Zusammenfassung

Das Volumen an massenspektrometrischen Daten wächst rapid dank neuer Akquisitionsmethoden und immer schneller werdende Massenspektrometer. Derzeit existiert eine Vielzahl von Programmen, die diese Datenmenge prozessieren können. Dennoch beschränken sich viele nur auf einen kleinen Teil der kompletten Analyse oder sind auf bestimmte experimentelle Fragestellungen zugeschnitten. Nur wenige Programme können alle Aspekte einer detaillierten Analyse bewältigen und gleichzeitig den Zugriff auf alle Daten ermöglichen, um weiterführende und spezialisierte Fragestellungen zu adressieren.

Die Massenspektrometrie hat sich in den letzten Jahren stark weiterentwickelt und der Schwerpunkt verlagerte sich im laufe der Zeit von der einfachen Identifizierung zur Quantifizierung einzelner Proteine. Tiefere Einblicke in die Auswirkungen von Krankheiten oder einer bestimmten Substanz auf die Proteinabundanzen in einem biologischen System oder das Verfolgen von Änderungen der Proteinmenge über die Zeit erfordern eine sehr akkurate und präzise Quantifizierung.

Ziel dieser Arbeit war es, ein Programm zu entwerfen und zu implementieren, das diese Entwicklung unterstützt und die erforderliche akkurate und präzise Quantifizierung ermöglicht. Es sollte einfach zu bedienen und vielseitig einsetzbar sein, sowie die Rohdaten und Ergebnisse der einzelnen Analyseschritte strukturiert zugänglich machen. Das Programm sollte auf bestehende Techniken aufbauen und diese, wenn notwendig, erweitern. Zugleich sollte das Programm als Open-Source-Lösung anderen Forschern zur Verfügung stehen, und somit helfen, Fragen zu beantworten, die bisher außer Reichweite waren.

Nach einer allgemeinen Einführung in das Thema der Massenspektrometrie und der Datenauswertung, beschäftigt sich der zweite Teil dieser Arbeit mit der Entwicklung des Programms isobarQuant. In diesem Teil werden die Struktur und die grundsätzlichen Aspekte und Funktionsweisen von isobarQuant erklärt. Es wird dazu detailliert in die Extraktion der erfassten Rohdaten, die Identifikation mit Mascot und die Quantifizierung mit einem selbst entwickelten Algorithmus eingegangen. Abschließend erörtert dieses Kapitel auch die Aggregation der identifizierten und quantifizierten Peptide zu Proteinen und die damit einhergehenden Probleme.

Der dritte Teil dieser Arbeit beschreibt die akkurate und präzise Quantifizierung, die durch isobarQuant ermöglicht wird. Hier werden die entwickelten Methoden beschrieben, die es mir erlaubt haben, einen Katalog mit mehr als 9.600 Proteinhalbwertszeiten in vier menschlichen und einem murinen Primärzelltyp zu erzeugen, welcher verwendet wurde, um die Langlebigkeit von Proteinkomplexen und die Variabilität zwischen den verschiedenen Zelltypen zu charakterisieren.

Im letzten Teil wird auf die Veränderung der Peptid-fragmentierung durch isobare Marker eingegangen und diskutiert, welchen Einfluss diese auf die Identifizierung durch Mascot haben. Zudem wird gezeigt, wie isobarQuant verwendet wurde, um eine neue Methode zur Aufdeckung potenziell co-eluierender Peptide und eine Verbesserung des H-Score-Algorithmus zu entwickeln.

Contents

List of Figures

Part I

Introduction & Objectives

1 Principles of data acquisition by mass spectrometry

In the last two decades, mass spectrometry (MS)-based proteomics has become a standard workhorse technology used in most core facilities at universities and research organizations throughout the world. The term ‘proteome’ was first used in 1995 and is attributed to Marc Wilkins to describe the full (or potential) complement of proteins expressed in a cell of the microorganism *Mycoplasma genitalium*,[1] where fewer than 100 proteins were extracted. Today it is routinely possible to identify in excess of 5000 proteins in a single, MS experiment, leading some scientists to argue that the global analysis of proteins and their function is far enough advanced to allow us to understand how cells function[2]. Mass spectrometry has of course been around for far longer. Thomson[3] and Aston[4] first used it in the determination of mass-to-charge ratios of electrons and later in the determination of isotopes of neon, chlorine and bromine. Throughout the 20th century its application was broadened, the pioneering work of McLafferty in the characterizing of chemical entities by relating spectra to structure[5] and even the first attempts at identitying amino acids using mass spectrometry[6]. It wasn’t, however, until methods to ionize peptides had matured and the availability of better and faster computational power and improved instrumentation became available that mass spectrometry started to revolutionize the proteomic space.

1.1 Mass spectrometry-based proteomics

Mass spectrometry-based proteomics can be roughly divided into two camps, depending on the desired throughput and biological questions to be answered. The first, known as top-down proteomics, aims to preserve whole proteins in their native state and environment, possibly even keeping protein complexes together. The second, bottom-up proteomics, works by firstly denaturing and then cutting the individual proteins into constituent peptides using a proteolytic enzyme. These protein pieces are then measured and re-assembled computationally to reveal the proteins in the original sample. Although the situation is slowly changing, the former is much less widespread within the proteomics community than the latter, which is predominant in proteomics labs throughout the world and was the setting in which this PhD was carried out. A brief description of both follows.

1.2 Top-down proteomics

Top-down proteomics has been around for many years primarily in the form of 2D-gel analyses, but top-down mass spectrometry (and a subset of it known as as native mass spectrometry, where proteins are analyzed in their ’native’ state), which enables the analysis of intact proteins and proteoforms using a mass spectrometer has recently been gaining some traction within the community. Its main limitation had been the requirement for high accuracy and high resolution instrumentation for the determination of the charge state of, for instance, a 20kDa protein which may carry up to 30 charges. It has been shown to be possible using an linear trap quadrupole (LTQ)-Orbitrap[7], but traditionally was the preserve of the fourier transform ion cyclotron resonance (FT-ICR) instruments (see section 3). The advantages mainly come from the ability to elucidate functions that would otherwise be missed by bottom-up proteomics (see next section 1.3). This includes detection of modifications that are removed or scrambled during peptide sample preparation (for example, S-thiolation[8]); the highlighting of functional relationships (for example, cross-talk) between post translational modification (PTM)s on the same protein molecule[9,10] and the identification and quantification of distinct proteoforms (proteins sharing large swathes of identical amino-acid sequence) that would have been convoluted by proteolytic digestion. Additionally, this bird's-eye view of the proteome is used in the determination of protein complex members[11] and in gaining insight into the stochiometry of proteins in the given sample. Despite these advantages and mainly because of the original requirement for highly specialized equipment, top-down mass spectrometry still inhabits a small niche relative to the much more popular bottom-up approach.

1.3 Bottom-up proteomics

Bottom-up proteomics concerns itself with the denatured and proteolytically-cleaved subunits of the proteome known as peptides. These peptides are generally more amenable to mass spectrometry than complete proteins, not only because measurement of their smaller size has a lower inherent absolute error, but also because they will have a much less complex mass distribution. Their smaller size also means they are much more readily solublized and separated. When cleaved with trypsin, the resultant peptides are highly suitable for mass spectrometry because the terminal arginine or lysine residues are protophilic and encourage good ionization of all peptides, which should contain at least one such residue. Routine analysis of entire proteomes is single mass spectrometry experiments is really only possible using a bottom up approach where the proteome is firstly denatured into proteins and then digesting it into peptides, separating these using chromatography (both online and offline) and then further breaking these peptides into fragments within the mass spectrometer. These peptide 'puzzle pieces' which are more easily handled can then be reconstituted following mass spectrometry analysis. It is this 'reconstitution' which was the primary focus of the work carried out during this project. A typical sample workflow from bench to mass spectrometric acquisition (following a biochemical or biological experiment and potentially a sample enrichment step) is described below.

protease name	cleavage specificity
Arg-C	C-term; arginine
Asp-N	N-term; aspartic acid
Chymotrypsin	C-Term; phenylalanine, leucine, tryptophan, tyrosine
Lys-C	C-Term; lysine
Glu-C	C-Term; aspartic acid, glutamic acid

Table (1): Table of enzymes and their amino acid specificity. In most cases the presence of proline in the subsequent position after the given residue will prevent cleavage

1.3.1 Sample preparation & separation by chromatography

1.3.1.1 Proteolytic digestion To break proteins into the smaller parts need for high-throughput mass spectrometry, proteases are used. A protease typically cleaves at the peptide bond within proteins. The most widely used for the mass spectrometry friendly reasons mentioned above, is trypsin which cleaves on the carboxy terminus of all arginine and lysine residues, unless the subsequent residue is proline[12]. Other proteases have different cleavage specificities and may be used instead of, or in concert with, trypsin to yield peptides that might otherwise not be suitable for mass spectrometry analysis. A selection of the commonest enzymes used in mass spectrometry experiments and their residue specificity is given in table 1.

The peptide mixture resulting from a proteolytic digestion of proteins in typical mass spectrometry experiment (containing many hundreds of proteins) is still highly complex and could not just be injected into a mass spectrometer. The peptides must be further separated based on their physico-chemical properties. This is usually achieved using liquid chromatography (LC), either with a direct feed into the mass spectrometer (online) or prior [and orthogonal to] the second, online, LC separation step (offline). There are several typical methods to separate peptides that are complementary to the online LC step and include strong anion exchange chromatography (SAX) or strong cation exchange chromatography (SCX)[13], isoelectric focusing (IEF)[14] or hydrophilic interaction chromatography (HILIC)[15]. It is at this stage that further enrichment and purification steps may be required to increase the abundance of a specifically desired type of peptide; for instance, peptides carrying PTMs may be of high biological interest, but exist at sub-stochiometric levels, leading to their being masked by other, more abundant signals. A number of different techniques have been developed to enrich for peptides containing particular PTMs: for example for phosphorylation[16,17], acetylation[18] or ubiquitination[19].

The high performance liquid chromatography (HPLC) system feeding directly into the mass spectrometer achieves separation of peptides by passing them through an immobilized porous substance with which the peptide will interact to varying degrees, depending on its physico-chemical properties. The porous substance is referred to as the stationary phase and the peptides (and the liquid they are transported in) is referred to as the mobile phase. Depending on how much each component interacts with the stationary phase, a different amount of time will be taken for it to pass through the system (the retention time). Ideally this time would always be same for an identical peptide. In HPLC the stationary phase is also referred to as packing and often consists of so-called C18 material present within a column of typically between 10 and 25 cms in length measuring approximately 30 mm in diameter. The '18' denotes that there are hydrocarbon chains with 18 carbon atoms bonded to the silica particles inside the column and it provides a very hydrophobic environment; the more hydrophobic the mobile phase (and the peptides) the longer they interact with the C18 material and hence the longer their retention time will be. During the course of the run the solvent of the mobile phase can be changed from less to more hydrophobic (from an aqueous acid to an organic solvent, for example), which changes the propensity of the peptides for the mobile phase from low to high, thereby altering (and decreasing) their retention time.

The eluate (fluid containing the peptides) coming off the LC column is at atmospheric pressure and feeds directly into the next destination on its journey: the mass spectrometer; but before that it must first be ionized.

1.3.2 Instrumentation and principles

Because a mass spectrometer fundamentally measures *mass-to-charge ratio (m/z)*, all analytes must be ionized (i.e. carry a charge) in order to be recorded. The ionization of the analytes is the first of the three main tasks performed by a mass spectrometer. Traditional ionization techniques in mass spectrometry included electron ionization and chemical ionization and were used for small molecules that are volatile and thermally stable[20], however, these harsh techniques were not suitable for the analysis of biomacromolecules such as proteins and DNA which do not readily vaporize. This problem was overcome during the 1980s and 1990s with the advent of so-called 'soft ionization' techniques like electrospray ionization (ESI)[21] and soft laser ionization, also known as matrix-assisted laser desorption/ionization (MALDI)[22] which do allow generation of gaseous ions from non-volatile macromolecules.

1.3.2.1 MALDI MALDI uses a laser energy absorbing matrix to create ions from large molecules with little or no in source fragmentation. MALDI is a three-step process. First, the sample is mixed with a suitable matrix material and applied to a metal plate. Second, a pulsed laser irradiates the sample, triggering ablation and desorption of the sample and matrix material. Finally, the analyte molecules are ionized by being protonated or deprotonated in the hot plume of ablated gases and are subsequently accelerated into the mass spectrometer.

1.3.2.2 ESI The ESI technique is the most common in proteomic mass spectrometry because it can be coupled directly to the HPLC where the flow of peptides at low concentration is directly infused into the mass spectrometer via a narrow needle. ESI works by applying a very high voltage between the needle and the detector orifice, which separates the charges at the surface of the electrolytic solution, causing the meniscus to deform into a cone. This Taylor cone forms as soon as a

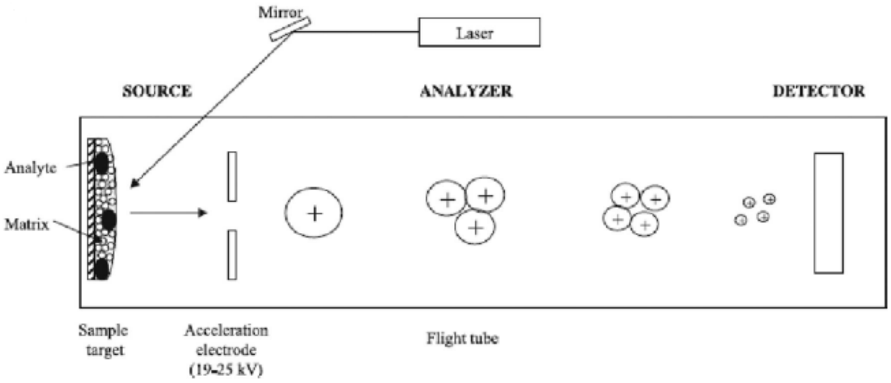


Figure (1): MALDI schematic taken from ref.[23] with permission. Sample is mixed with matrix material and applied to a plate. This is then irradiated by a laser to ablate the sample whose molecules are ionized and accelerated towards to the mass detector.

critical electric field is reached and leads to the ejection of a jet of liquid ions of the same charge sign originating at its apex. This aerosol jet is unstable and breaks apart into small droplets from which the solvent progressively evaporates, leading to yet smaller droplets. Once the droplets reach the Rayleigh limit they explode into a spray of charged ions heading towards the detector entrance. This is illustrated in figure 2. The use of an even smaller needle is given the term nanoESI. Typically ESI produces ions with a charge of +2 and above while MALDI usually yields ions with a single positive charge.

The ionized analyte is now subject to the second task of the mass spectrometer: it is separated out according to the *m/z* of the constituent ions. These ions (their *m/z* and intensity) are recorded by a detector within the analyzer and computationally conduced into a mass spectrum. There are five different kinds of mass analyzers, grouped according to the mode they operate in. Hybrid instruments, as the name suggests, have more than one in-built analyzer and take advantage of the different benefits of each type of analyzer. The mass analyzer's primary function, as stated above, is to separate the peptides based on their *m/z*. They may enable the selection of an appropriate *m/z* range for subsequent fragmentation, and can also perform the fragmentation and measurement (detection) of the resulting fragment ions themselves. Peptide fragmentation and tandem mass spectrometry is discussed in greater detail in the following section (1.3.3).

1. time-of-flight (TOF) analyzers. Here the analyzer is a chamber under vacuum that itself contains no electric fields. Following ioniziation, the peptides are accelerated by an electric field into the analyzer. The ions drift through the analyzer with the kinetic energy obtained from the potential energy of the electric field. Assuming that all ions obtain the same energy, the ions of lesser *m/z* will have greater velocity than ions of greater *m/z*. Therefore, as ions traverse the analyzer, they separate out according to their *m/z*. A detector is positioned at the end of the analyzer to measure the arrival time of ions. TOF analyzers are often, but not always coupled to MALDI ion sources. The quadrupole time-of-flight (QTOF) instrument is coupled to a standard LC system with an ESI

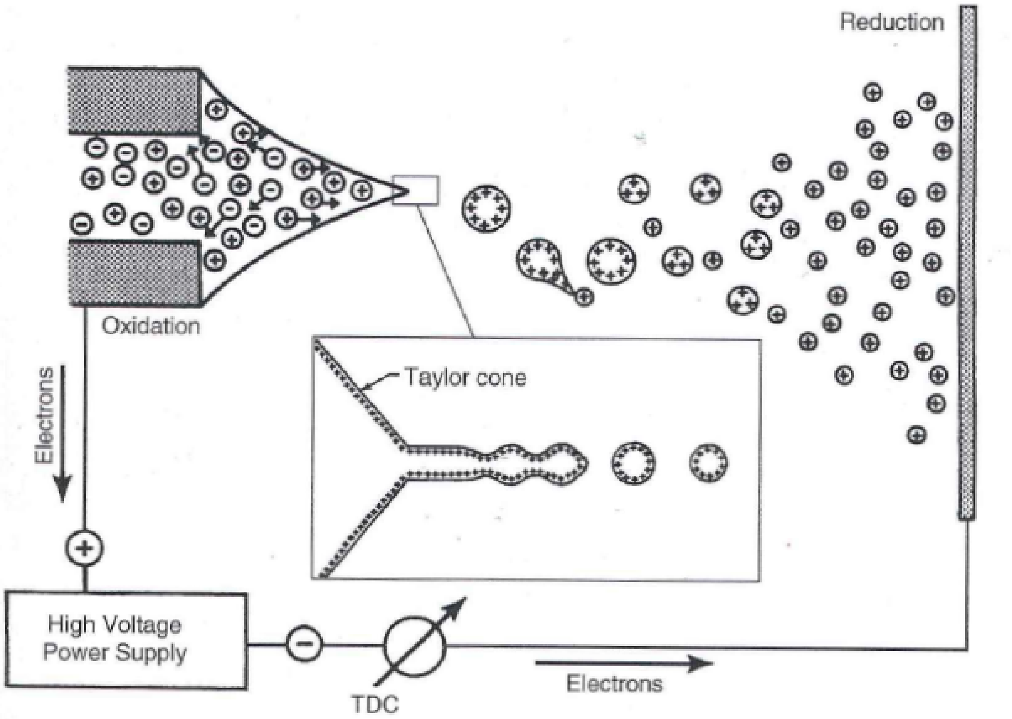


Figure (2): Taylor cone and electro-spray formation taken from ref.[24] with permission. The cone forms because of charge separation of the electrolytic solution which in turn deforms its meniscus. When a critical electric field is reached an aerosol jet is formed. This unstable jet breaks apart into ever smaller droplets as the solvent evaporates which then head towards the entrance of the detector

source[25]. A recent develpment in theTOF technology sphere comes with the introduction of the trapped ion mobility spectrometry (TIMS)-TOF instrumentation where the ionization source is nanoESI[26] and offers additional information about the three dimensional structure of ions being acquired, .

2. Quadrupole analyzer. A quadrupole mass analyzer consists of four metal rods arranged in parallel to which direct current and radio-frequency voltages are applied. Gas-phase ions entering the mass analyzer follow a corkscrew trajectory along the axis of the quadrupole, the radius of this trajectory depends on the *m/z* of the ion and the offset voltage for the field. The voltage applied can therefore be used to select the *m/z* range of the ions that are allowed to pass through the quadrupole region. All ions outside this range will not go through and hence will not be detected. Ions of increasing *m/z* can be detected by sweeping the radio frequency voltages on the quadrupoles.

3. The FT-ICR analyzer. This analyzer contains a cyclotron that accelerates particles to high energies. The ions rotate around a strong magnetic field (up to 9 Tesla) and when an electric field oscillating at, or near, the cyclotron frequency of the trapped ions is applied, it excites the ions into a larger orbit where they may be measured as they pass detector plates on opposite sides of the trap, or expelled if outside the required *m/z*. Their cyclotron frequency is then transformed using a Fourier transform to yield the *m/z* and intensity values. The resolution and accuracy of this type of analyzer is extremely high.

4. Ion trap analyzers. As the name suggests these analyzers trap ions in a confined space. They can be divided into three groups: the 3D ion, (or quadrupole ion-QIT-, or Paul-) trap, the linear ion-, linear ion trap (LIT)- (or Penning-) trap and the Orbitrap, which is covered in the section below. The QIT is like a quadrupole (above) but with two of the rods forming the endcap electrodes, the third is formed into a ring and the fourth is collapsed to the middle. The ions are captured by being alternately compressed and expanded along the *x*-axis (which runs from the source to the detector). Ions enter from the source through one of the endcaps and the *m/z* of interest is selected by applying the correct voltage while all other ions are ejected. A collision gas is applied to the ions and their kinetic energy is increased, which leads to fragmentation of the parent ions into smaller ions. These smaller, fragment ions can then be read out sequentially by scanning through a range of voltages at the detector. quadrupole ion trap (QIT)s have poor resolving power and suffer from poor accuracy and have a limited dynamic mass range. They are however quite sensitive. LITs improve on QITs by having a simpler construction: four parallel electrodes which trap a larger number of ions.

5. Orbitrap: The idea of the Orbitrap mass analyzer was conceived in the year 2000 by Alexander Makarov[27] and was later published by his group at Purdue University in 2005[28] and represented the first high resolution, high mass-accuracy mass spectrometer at a price affordable by a standard proteomics laboratory[29]. It is ultimately based on an ion trap design dating back to the 1920s by Kingdon[30] and then in the 1980s by Knight[31]with the main principle of the Orbitrap focusing on trapping ions in a constant radial electrostatic field generated by an outer electrode. Makarov exploited the oscillating motion of the ions along the central electrode. The *m/z* value of the ions is calculated using a Fourier Transform, based on the frequency with which they oscillate along the central spindle. The Orbitrap is similar in resolution and accuracy to FT-ICR instrumentation but does not require any magnets to hold the ions and has no radio frequency to initiate their motion.

1.3.2.3 Hybrid instrumentation It makes logical sense to try combine the advantages of the different types of analyzers and ion guiding devices into one instrument depending on the aims of the manufacturer whether that be highest resolution and accuracy, improved speed and accuracy or creation of a compact, versatile and affordable machine. Usually the most important driver is the cost which, in the case of tandem mass spectrometry translates to having a lower resolution first stage (MS1) mass analyzers, with a high resolution second analyzer (for MS2) to measure product ions.

The source of all of the data that was used for the development of the methods presented in this PhD thesis has been from the Orbitrap-hybrid family of instrumentation.

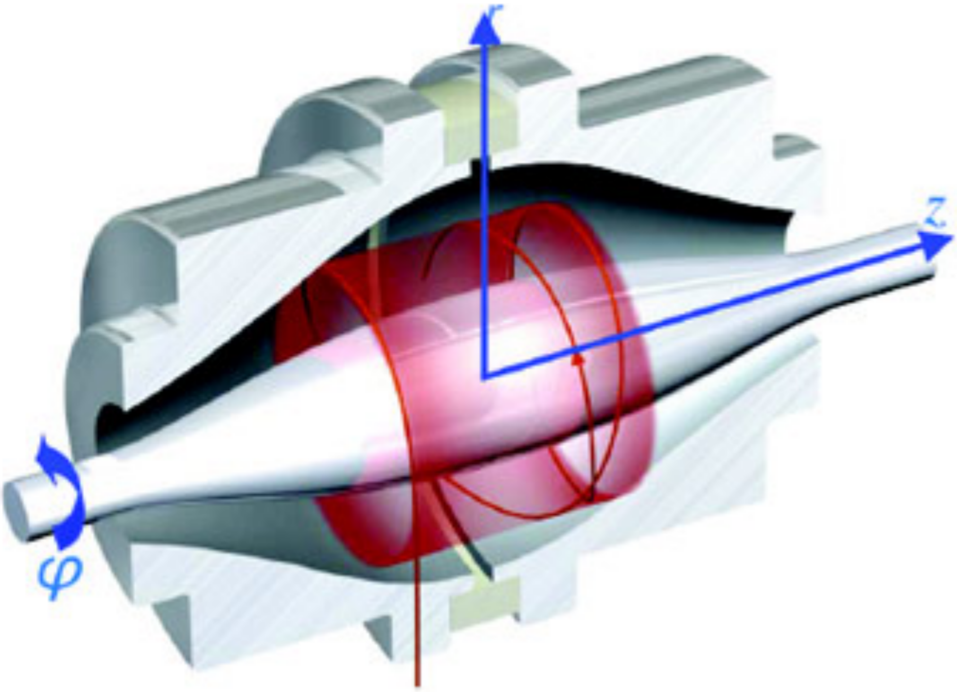
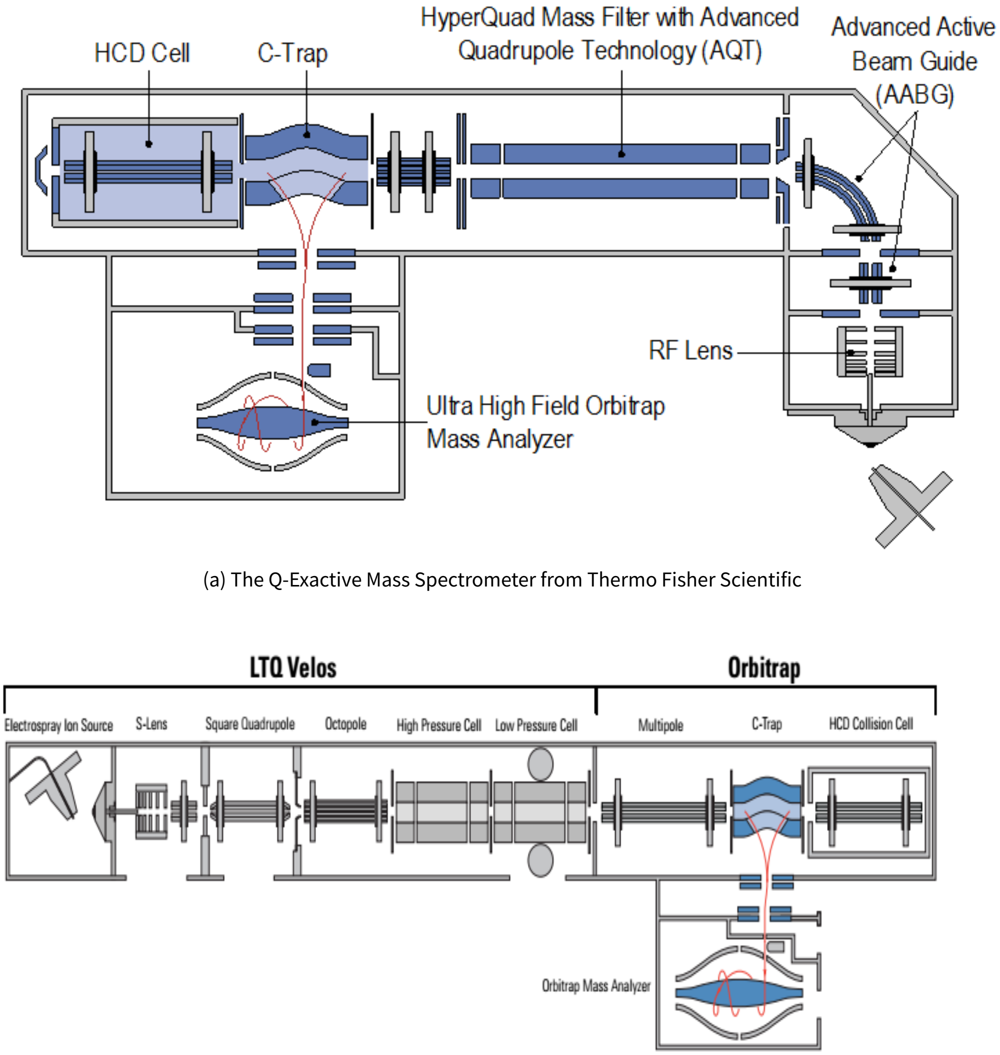


Figure (3): Depiction of the Thermo Fisher Scientific Orbitrap (taken from ref.[24] with permission). Ions move in a spiral fashion around the spindle-shaped central electrode and separate out along the axis according to their *m/z*. The *m/z* of the different ions can be determined from their frequencies of oscillation following a Fourier transform

The two main instruments featured in this report are the Q-Exactive, which was released in 2011[32] and built on the successes of the earlier Orbitrap instruments. The other workhorse instrument which provided data for this project is the Orbitrap Velos[33] which was released in 2009. The main improvements between the two were in terms of scan speed / cycle time and in the resolution of the Orbitrap. The Q-Exactive operates exclusively in higher energy dissociation (HCD) mode (acquiring MS Scans and MS/MS spectra in the Orbitrap). The Velos is able to use either Orbitrap or the ion trap for the acquisition of MS/MS spectra (in HCD or collision induced dissociation (CID) modes respectively).

1.3.3 Tandem Mass Spectrometry Experiments

Analyzing and recording the *m/z* of intact ions eluting from the LC column can provide a good deal of information about the analytes present in the sample. However this information may not be enough to unambiguously map the recorded ion onto a specific peptide sequence. This is particularly true with complex samples where several different peptide sequences occur within the tolerance (or resolution) of the mass spectrometer for a single given ion or when modification of one or more of the peptide's amino acid residues alters the expected mass. In order to correctly sequence the peptide, further analysis steps are necessary. Firstly, the mass analyzer is instructed to select *m/z* of ions of interest until a minimum abundance threshold is met. The collected ions corresponding to the desired *m/z* are then fragmented into smaller ions. The resultant *m/z's* and intensities of the fragment ions are recorded by the detector via one of the ways described for the first step. This second step, where



(a) The Q-Exactive Mass Spectrometer from Thermo Fisher Scientific

(b) The LTQ Velos Mass Spectrometer from Thermo Fisher Scientific

Figure (4): Schematic of two of the workhorse instruments featured in this report taken from ref.[34]

the sequence information is obtained, is referred to as Tandem MS or MS/MS with the second, fragment ion spectrum being referred to as the MS/MS or the MS2 spectrum. The first and the second MS acquisition may be separated in either space (a different analyzer is used to acquire the second spectrum) or in space (in that analysis occurs in the same analyzer but subsequent to the first acquisition). It is possible to continue the cycle of fragmentation and analysis of generated products for further rounds of fragmentation for as long a time as there are sufficient ions to collect and fragment. This kind of acquisition is referred to as MS^n. Often fragmentation is halted after the third iteration and, combined with the synchronous precursor selection on specific instrument types, has been shown to mitigate the problem of co-isolating peptides in isobaric quantification[35–37] (see section 1.4.2.3).

1.3.4 Peptide fragmentation

In most cases, peptide fragmentation occurs along the peptide backbone and results in regular fragments along the length of the peptide. If the fragment is from the N-terminal part of the peptide it will be referred to as an a-, b- or c-ion, depending which peptide bond is broken. If the fragment is from the C-terminal part of the peptide it is the named x-, y-, or z-ion. A subscript notation will denote the residue of the peptide. An example is shown in figure 5. Apart from these backbone fragments it is possible to find internal fragments, where double backbone (or peptide) fragmentation has occurred and immonium fragments, where a single residue side chain is identified surrounded by an a-type and a y-type fragmentation. This standard naming convention was established by Roepstroff and colleagues[38]. Fragmentation can occur spontaneously (and simultaneously) during the ionization step, but this is not usually desired in mass spectrometry proteomics and is another reason for choosing the soft ionization technique. The main means to perform the fragmentation is in a dedicated, separate part of the mass spectrometer and there are several methods to do this, which are discussed below.

1.3.4.1 CID CID or collision associated dissociation (CAD) is performed in a dedicated part of the mass spectrometer, referred to as a collision cell. Upon entry into the collision cell the ions are accelerated with an electrical potential, which increases the ions' kinetic energy. This leads to an increase in the number of collisions with an inert gas such as helium or nitrogen which has been introduced to the cell. Following a collision the kinetic energy is converted to internal energy and initiates bond disruption. The resultant fragments have a lower *m/z* and are no longer excited by the applied electric potential and are thus not further fragmented.[39] This phenomenon may lead to problems in the analysis of labile modifications, such as phosphorylation because the weak bonds are preferentially broken. CID generates predominantly b-, and y-type ions. CID performed in a quadrupole ion trap typically suffers from the low *m/z* cut off which would normally render it unsuitable for MS2 based isobaric labeling (see section 1.4.2.3). A workaround for this was found by Schwarz and co-workers, termed Pulsed Q Collision Induced Dissociation (PQD)[40,41], however this has largely been superseded by either HCD fragmentation[42] which does have this limitation or by using multinotch MS3[35]. In Triplequad (QqQ) instruments the collision cell is typically the second quadrupole.

1.3.4.2 HCD HCD [42] is also known as is high energy CID or beam-type CID or originally higher-energy C-trap dissociation works in more or less the same way as CID but with higher collision energies. It is performed in a dedicated collision cell and works with high resolution ion detection, and increased ion fragments with no low mass cut off. In spite of its name, HCD is still referred to as a 'low energy' collision induced dissociation (less than 100 eV)[43]. HCD and yields not only b- and y-type ions, but is also rich in a-ions and immonium and internal fragments[44]. This large range of different fragment ions types is brought about because the higher energy allows fragment ions to further decay into smaller components. This fact and the lack of a low mass cut off renders HCD to be highly suited to MS2-based quantification strategies.

1.3.4.3 ECD electron capture dissociation (ECD)[45] typically involves a multiply-protonated molecule interacting with a free electron to form an odd-electron, radical, ion. Liberation of the electric potential energy (neutralization of one ionic charge) results in fragmentation of the product ion. This method of fragmentation often requires additional hardware to standard instrumentation set up. It is primarily used to generate c- and z-type ions complementary to the typical b- and y-ions produced by CID (HCD). The nature of this type of fragmentation preserves labile PTMs such as phosphorylation and glycosylation and can be useful in PTM site localization.

1.3.4.4 ETD electron transfer dissociation (ETD)[46] also creates predominantly c- and z- type ions and also keeps modifications and amino acid side chains intact. It does not have the requirement of ECD of an ultra-high vacuum and is amiable to LIT instrumentation. The electron donor reagent anions required for the creation of the radicals is typically fluoroanthene obtained by external chemical ionization source which is injected into the center of the LIT and then mixes with the protonated analytes. This method generally requires higher charge states than the typical +2 ions following tryptic proteolysis so a different enzyme might be used. As for ECD this method is complementary to CID and may require additional instrument hardware.

An important aspect of peptide fragmentation is that the ions collected by the analyzer for fragmentation during the first acquisition will not necessarily be of one exact *m/z* but rather of a range of different *m/z*'s within a window, and so may consequently contain more then one peptide. Thus the fragment ions may contain the products of fragmentation of more than one parent ion. The abundance (intensities) of the different product or fragment ions detected and their role in determining the peptide sequence of the precursor ions selected by the MS1 analyses lies at the heart of many of the algorithms described later on (section 1.5.1). and is a fundamental aspect of mass spectrometry-based proteomics.

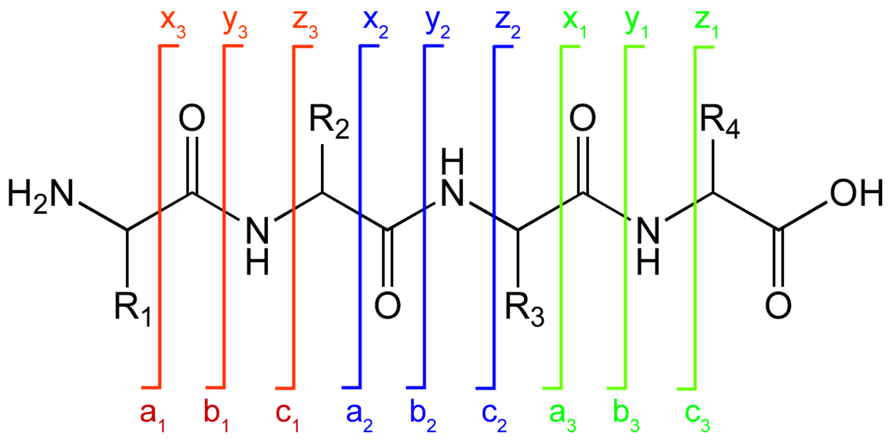


Figure (5): Example of peptide fragment notation (taken from ref.[47]). a-, b- and c- fragments are formed if the charge is retained on the N-terminus and x-, y- and z- fragments formed if the charge lies on the C-terminus. The subscript number corresponds to the position of the break counting from the respective terminus.

1.3.5 Two modes of peptide fragmentation

As sketched out above, proteomic mass spectrometry works by first scanning and measuring ionized entities eluting from the LC column, selecting an *m/z* range and then fragmenting the corresponding ions and subsequently measuring the product *m/z*'s. There are two main modes of acquisition to obtain the product ions: Data dependent and data independent acquisition. As the name suggests, data dependent acquisition depends on the first MS1 scan to find the most abundant ion species and then collects ions corresponding to the given *m/z* range and subsequently fragment them, while data-independent acquisition, which is independent of the MS1 ion intensities, simply divides up the MS1 scan into regions of *m/z* and sequentially fragments all ions present for each given region.

1.3.5.1 DDA data dependent acquisition (DDA) mode is also known as 'shot gun mode' because of its inherent stochastic nature which is likened to the 'quasi-random firing pattern of a shotgun'. It is currently the most frequently used acquisition mode in proteomic mass spectrometry. Owing to the fact that this mode records and uses precursor ions of a known *m/z*, this value is available to the downstream analyses to limit the number of potential peptides matching to the derived fragment ions (see section on fragment matching and peptide search engines 1.5.1). There are a number of other DDA modes which operate at the MS1-scan level, but which use a predefined list of masses to fragment within a given retention time window. These targeted modes may also include further information (transitions) to select particular fragment ions. These can be particularly useful when performing quantification of analytes in a label free manner since it is possible to be certain, even in independent runs, that the intensity of the same peptide and fragment are being measured. The individual signals can be integrated over the LC run and if required compared between runs to give a ratio of abundance in two ore more conditions.

1.3.5.2 DIA Data independent acquisition (DIA) describes the instrument set up when a set of constant mass ranges independent of the peptides analyzed in the MS1 scan is isolated for fragmentation. This method presents quite a number of challenges with regard to the development of robust data analysis tools. The multiple peptides in each *m/z* window are fragmented resulting in complex MS/MS spectra which require difficult deconvolution. Essentially, all MS/MS spectra from all MS1 ions are acquired irrespective of abundance in the first scan. Having this fully comprehensive and correspondingly large amount of data may offer some advantages over DDA in situations where large sample cohorts are to be quantitatively compared, for example in exploratory analyses looking at well-known sample types such as in plasma or urine. It has also been claimed that DIA can identify more peptide to spectrum match (PSM)s than the theoretical number of MS/MS spectra that can be sequentially acquired in a DDA run, because nothing is missed out[48], this also means that different questions can be asked of the data and computationally analyzed as many times as required until an answer is found. DIA has shown a lot of potential and is already generating some interesting results. For now, this method remains outside the scope of this PhD thesis and will not be discussed further here.

1.3.5.3 Targeted A third mode, Targeted (Data Dependent) Acquisition, which may be categorized as a subset of DDA and is fundamentally very similar to it. Following the MS1 scan, only peptide *m/z*'s identified within the scan and present on a user-defined 'wish-list' within the given matching retention time window are selected for fragmentation. This method aims to enrich for lower-abundant precursors that might not otherwise be selected for fragmentation. It is a way to circumvent the problems associated with the wide range of peptide abundances eluting at the same time from the LC column. Several different targeted approach (TA) methods exist depending on the type of instrumentation in use. multiple reaction monitoring (MRM) and selected reaction monitoring (SRM) are very similar and are largely performed on triple quad instruments and parallel reaction monitoring (PRM) on (mainly) trap-types of instrumentation.

The three methods described are all summarized in figure 6.

1.4 Methods for quantifying proteins

In its first twenty years, proteomics was a qualitative science whose primary focus was on reconstructing a catalog of proteins present in a sample at the time of acquisition. In the last decade mass spectrometry-based proteomics has become more powerful thanks to the addition of a further dimension: quantification. It is now not only possible to say whether a protein is present or absent in a biological system but also by how much. To a certain degree this had been possible previously by counting the number of PSMs for one protein in a single experimental run and comparing it to

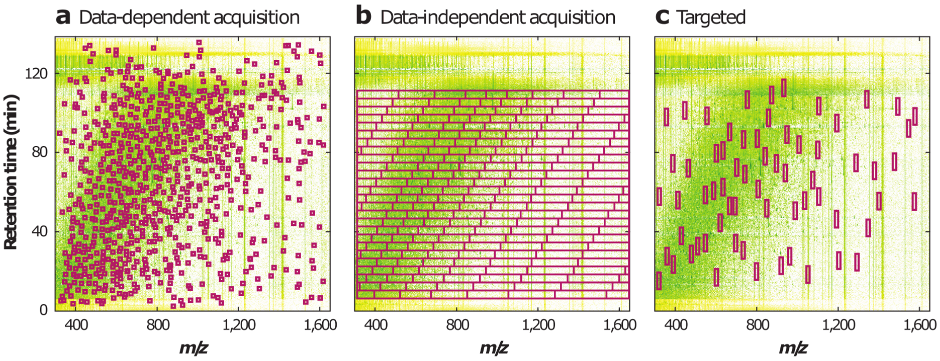


Figure (6): Three modes of selecting analytes for fragmentation. Figure taken from ref.[49] with permission. The red squares indicate the *m/z* region selected for fragmentation. Panel (a) describes data-dependent acquisition, where, over time, narrow precursor windows are selected based on the intensity within the MS1 scan and a time-limited exclusion list of precursors preivously fragmented. Panel (b) illustrates a data-independent approach where the entire *m/z* range is sampled via wide *m/z* windows over a longer retention time duration. Panel (c) shows a targeted acquisition approach where, over specific retention times, a narrow *m/z* range (containing a peptide of interest) is selected for fragmentation. This aim to mitigate problems associated with the large dynamic range of peptides eluting from the LC column.

the number for the same protein in a different run but this was subject to a battery of problems and issues.

Several new techniques and approaches coupled with improvements in instrumentation and LC set-ups have helped bring quantification of proteins by mass spectrometry into a routine operation. Quantification in proteomics has two approaches: absolute and relative quantification. The former aims to establish true amounts (as in copy number per cell or actual concentration) of a given protein within a sample, while the latter aims to uncover the ratio or change in protein abundances between two or more samples.

1.4.1 Absolute quantification

In the DDA approach of bottom up proteomics, absolute protein quantification is typically achieved by obtaining a regression function between the raw signal generated in an experiment and a known quantity of a spiked-in protein. Silva and colleagues observed that the top three most intense peptides remained constant within a coefficient of variation of +/- 10%[50] and used this value to estimate absolute protein abundance in a sample (see also section 2.3.3.1). A very similar approach, named intensity-based absolute quantification (iBAQ), was taken by Schwannhäusser *et al.*[51] to use the intensity of all peptides associated to a protein, but to normalize these by the number of possible MS-visible peptides. An alternative approach; the proteomic 'ruler' uses the MS1 signal from histone proteins (proportional to the amount of total DNA), which is dependent on the number of cells to estimate the copy number of a protein per cell[52].

1.4.2 Relative quantification

Performing relative quantification is easier than absolute quantification due mainly to how the instruments detect ions. This means the read-out of a such a quantification experiment is a fold change in abundance across samples. This kind of quantification can be further divided into two types: label free and labeled, where the label refers to the addition of a chemical label to induce a mass shift which distinguishes peptides originating from different conditions. Label free aims to estimate the change in peptide abundance from different mass spectrometry runs, usually by assessing the relative precursor intensities.

1.4.2.1 Label-free quantification The label free approach is in essence the simplest: One directly compares the abundance of peptides in one sample run with the abundance of the same peptides in one or an infinite number of other sample runs. The relative abundance may be estimated using methods such as spectral counting[53], where one counts the number of MS/MS events triggered for a given peptide; spectral counting normalized by protein length or number of observable peptides[54], which is the same as the previous approach but tries to normalize for differences in protein length; or by the integration of ion intensities (either precursor or fragment ions) over the chromatographic profile of the run(s) analyzed[55]. All of these methods require multiple (sequential) acquisitions and highly reproducible sample preparation and handling. Although ions observed in an MS1 scan are generally proportional to peptide abundance in the the given sample(s), absolute signal intensities can vary depending on a number of factors such as small differences in the chromatography, slight changes in instrument performance and variations in peptide ionization efficiency. This leads to high variability and increased numbers of missing data points from one run to the next.

This between-run variability can be mitigated by combining and measuring all samples together into a single run. Here the addition of stable isotopes of elements such as ^{13}C, ^{15}N, ^{18}O and ^{2}H, incorporated into the samples in labeling steps upstream of the acquisition, allow the peptides from the separate starting conditions to be differenciated. The peptides behave largely identically in terms of their chemical and liquid chromatography properties, differing only in their mass or in a series of their fragment ions. This enables the corresponding intensities to be distinguished in the same MS1 or MS2 scan and quantified relative to one another or a single channel.

1.4.2.2 MS1 labeling Among the first MS1 labeling developments was isotope-coded affinity tags (ICAT). Here heavy and light peptide alternatives are created through biotin affinity tags linked to the thiol reactive group of cysteine-containing peptides[56]. Despite its success, this approach was clearly limited because cysteine is comparatively rare leading to a large number of peptides going unlabeled; alternative chemical labeling methods, based on ^{18}O[57] and dimethyl labeling[58,59] then followed suit. These chemical labeling approaches could, however, only account for differences introduced after the labeling step and deuterated peptides were subject to small shifts in their chromatographic profiles. This led to the development of metabolic labeling methods which aimed at reducing variability introduced during sample preparation. One example of this is stable isotope labeled amino acids in culture (SILAC)[60] where in one biological system naturally occurring amino acids (arginine and lysine) are entirely replaced by synthetic versions fully labeled with^{13}C and^{15}N isotopes. These ‘heavy’ labeled peptides are compared to naturally occurring ‘light’ amino acids to yield a duplex quantification method, which can be extended to triplex mode by reducing the numbers of incorporated isotopes from fully labeled $^{13}C_6{}^{15}N_2$-lysine and $^{13}C_6{}^{15}N_4$-arginine to $^{13}C_6$-lysine, $^{13}C_6$-arginine and $^{15}N_4$-arginine residues. In theory it should be possible to extend SILAC to a 5-plex mode but overlaps in the isotopic distributions of the different labels start to present limitations.

A truly metabolic strategy, where full replacement of naturally occurring nitrogen atoms with ^{15}N isotopes is another MS1-labeling alternative. This is used in plants and some higher eukaryotes where the heavy nitrogen is supplied in the culture medium. However, since the mass shift introduced by ^{15}N labeling affects the amino acid composition of all residues in the peptide (to differing degrees), knowledge of the peptide sequence is essential to calculate the expected mass difference between a labeled and unlabeled partner, which complicates subsequent data analysis. This is one likely reason why ^{15}N metabolic labeling is not so frequently used in proteomic experiments.

The Super-SILAC[61] approach is an extension to SILAC which aims to provide additional ’plexing capability by using a set of synthetic, isotopically labeled standard references in a series of binary comparisons across a large set of experiments. Each sample’s comparison to the common reference yields a ratio. The ratios of all the references can then be used to normalize the ratios to then find the overall ratio between the different samples. Drawbacks include the need for multiple acquisitions (and the associated variability); considerable bioinformatic analysis time; under-sampling of the proteome and the extra analysis time spent examining and fragmenting redundant precursors from the reference in the acquisition of every mix. Such a ratio of ratios approach will be affected by both the abundance of the reference and the measured sample(s) themselves.

Because of its early combination of samples, SILAC is the most accurate quantitative MS method currently available, which makes it suitable for assessment of relatively small changes in peptide and protein amounts. A discussion around this and a new computation software method to analyze these samples is presented in part III of this thesis.

A metabolic labeling method which promises to increase the multiplexing capacity for MS1 labeling was introduced by the Coon lab in 2013. Neucode[62] utilizes the neutron mass deficit (the fact that nuclear binding energy is different for isotopes of different elements[63], resulting in tiny mass shifts) to build labels with millidalton (mDa) mass differences. The distinct backbone fragments can be distinguished from each other in a dedicated high resolution (480, 000 @ 400 *m/z*) quantification scan following a typical MS1 scan performed at standard resolution, where these differences remain undetectable, resulting in reduced sample complexity with only a sin-

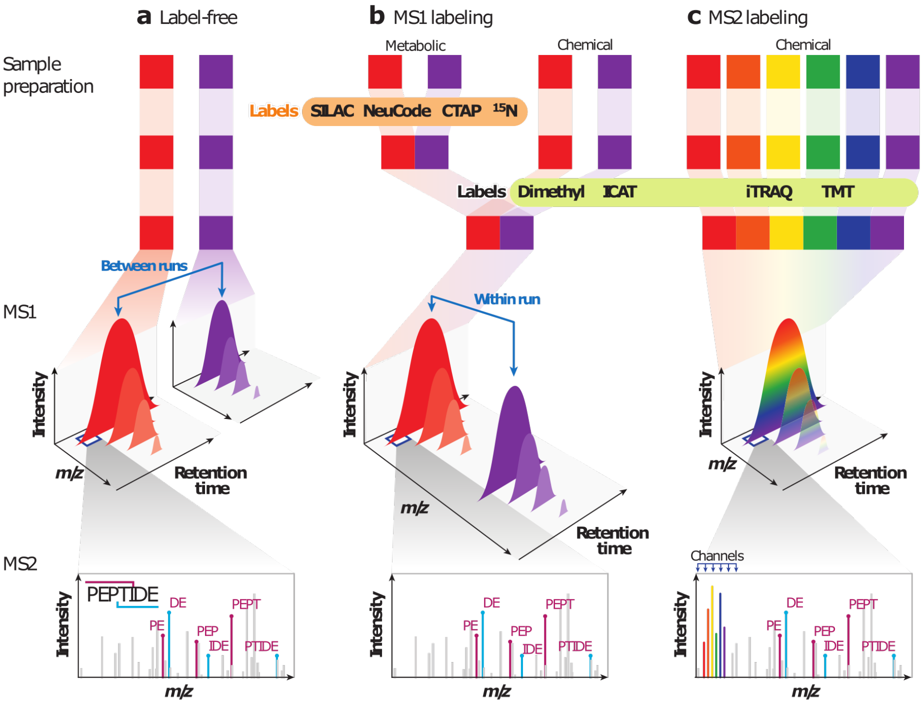


Figure (7): Three different methods of relative quantification taken from ref.[49] with permission. Panel (a) highlights a label free method where some measure of the peptide abundance is compared between two separate acquisitions. Panel (b) shows an MS1-based quantification approach where, within a single mass spectrometry acquisition run, the abundance of two peptide species differing by a known mass shift is compared. This mass shift may be introduced by a number of different methods. Panel (c) exemplifies an isobaric mass tag approach where, within a single tandem mass spectrum, the abundance of several different peptide species may be determined by reporter ions of known masses resulting from fragmentation of the precursor peptide.

gle tandem spectrum from two or more parent ions. The extra scan takes longer and the procedure can only be performed on high resolution mass spectrometers and would be beholden to the same problems as other MS1 methods.

1.4.2.3 MS2 labeling The second major approach to perform quantification in mass spectrometry is via isobaric labeling of peptides. Here a chemical reagent is covalently bound to peptides at the N-terminus and (usually) at the amine group of lysine residues. These labeled peptides are not only identical in terms of mass but also regarding their physico-chemical properties and behavior during separation on the liquid chromatography column. This results in all conditions being represented by a single MS1 feature: their differences are only revealed upon fragmentation. At this point the tag dissociates into its two component parts: a charged reporter ion and the balancer group (usually still attached to the peptide). Thanks to the arrangement of stable isotopes of nitrogen and carbon within the tag, the reporter ions will differ slightly in mass and can then be used for relative quantification. The two main methods using this approach are isobaric tag for relative and absolute quantification (iTRAQ)[64] available in a four and eight-plex capability and TMT[65,66] which, using the classic TMT moiety, encodes up to eleven different samples. During the writing of this thesis a further TMT-capability with a different structure was announced at ASMS 2019 (https://www.asms.org/conferences/past-conference). This TMTpro tag increases the number of possible quantification channels to sixteen. The chemical reagents are typically small (< 310 Da) and result in fragment ’reporter ions’ in the region of the MS/MS spectrum between 117 and 134 Da (inclusive) and are added to samples following (tryptic) digestion and prior to the samples being mixed and measured. Whilst the method offers many benefits over MS1 methods (the labels can be added to any sample irrespective of source, which is not possible with metabolic labels, and allows comparison of a higher number of conditions in one experiment with few missing values), it can suffer when the reporter ions in the MS/MS spectrum originate from one or more different precursor peptides. This phenomenon, dubbed ’ratio compression’, was described by Bantscheff *et al.*[67] and separately by Ow *et al.*[68], in the context of iTRAQ isobaric labeling. Several solutions have been put forward for this problem, which is discussed in more detail in the Chapter II, section 1.3. A relatively new extension to the idea of isobaric tagging (in part to circumvent the problems of ratio compression) was published by Winter *et al.*[69]. With the EASI-tag solution, the balancer group attached to the peptide-fragment (which is therefore specific for the given peptide) is used for quantification rather than the low-mass reporter ions which are devoid of any peptide information. This could potentially completely overcome the problems associated with ratio compression as only peptide-specifc reporter ions are quantified. However, the resolution at the higher *m/z* is lower, meaning that for a set of TMT-labeled samples, only six quantification modes would be possible.

All quantification methods have their associated advantages and disadvantages. The decision regarding which method of quantification should be based on the (biological or medical) question being asked; the availability and type of sample and the number of conditions to be compared. To a large extent, this thesis focuses on providing a software solution for the processing of data to accomplish both absolute and relative quantification of MS1 and also MS2 types.

1.5 Bioinformatic and computational methods employed in the processing of mass spectrometry data

The nature of the data coming off a mass spectrometer means that very little meaningful interpretation is possible without the intervention of a number of computational processing steps, this is compounded by the volume of the data recorded in each acquisition. Interrogating this would not be possible without computational support. Here I describe some of the methods required on the journey from the acquired, raw spectrum to the peptide and later, protein with the possible addition of post translation modifications and quantification. One could view this as the exact reverse of the journey that the sample took from protein to peptide to spectrum. This voyage-in-reverse aims to put the pieces back together to view a snapshot of the conditions in a cell or set of cells at a given time point under a specific condition.

1.5.1 Extracting information from tandem MS spectra and peptide searching

1.5.1.1 From spectrum to peptide The result of most DDA mass spectrometry experiments is a list of precursor masses and a tandem MS/MS spectrum corresponding to the *m/z* of the ions yielded by fragmentation of the given precursor and their corresponding intensities. These fragment ions can be used in conjunction with the precursor mass to provide evidence for the composition of the peptide that was present in the original mix. For a given mass in a complex peptide mixture, the number of potential matches to proteolytic peptides in a large proteome is usually is too high to rely solely on a peptide mass fingerprinting method[70–72] (where a single precursor mass maps onto a single peptide sequence within the mass tolerance of the mass spectrometer), hence it is necessary to include the masses of the fragment ions to increase the chance of a correct match. This combined effort (precursor and fragment ions) is the basis for the spectral matching methods presented here and can be divided into three main categories:

1. Spectra are matched to peptides generated from a protein sequence database which is digested *in silico* and the fragment ions series are calculated.
2. Spectra are matched to a library of pre-measured, validated, consensus MS2 spectra.
3. *De novo* and machine learning approaches.

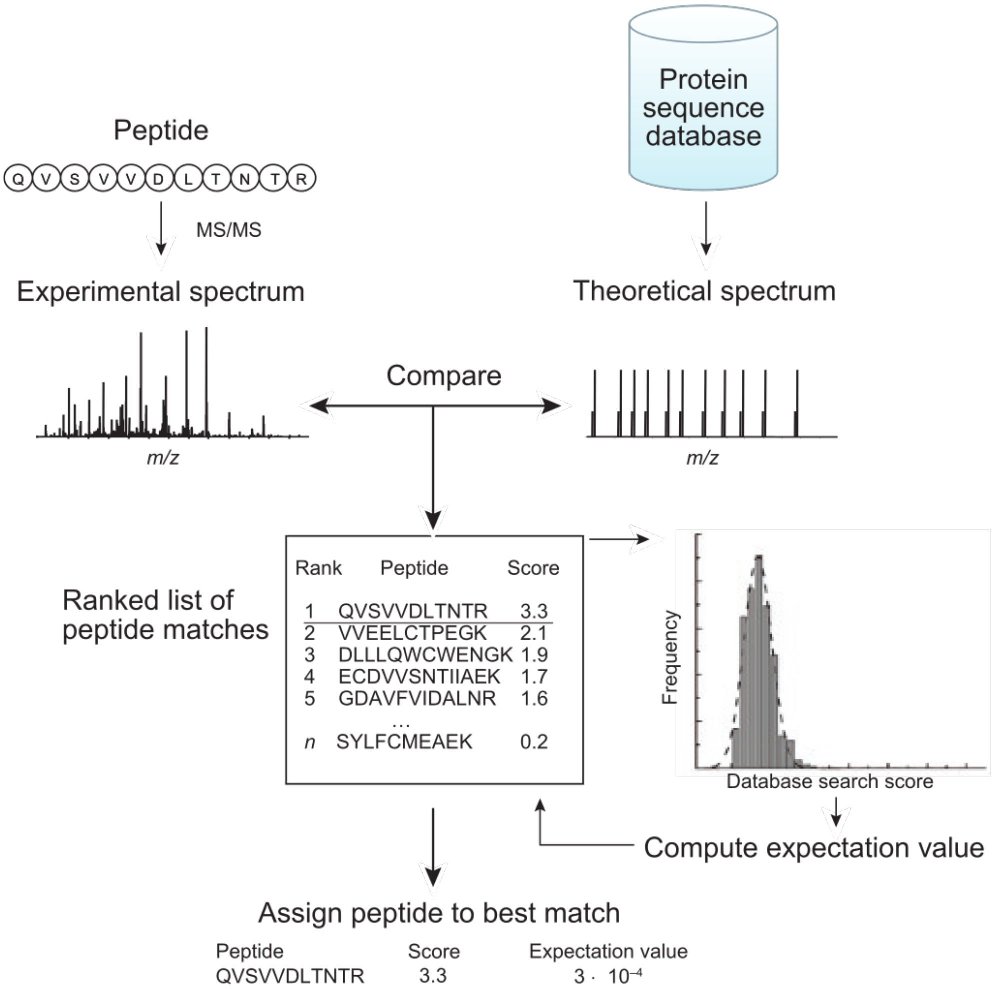
A fourth category consists of methods which combining aspects of each of the above.

1.5.1.2 Database searching The most widely used spectrum matching method is sequence database searching. Here an experimental tandem mass spectrum is compared to a set of theoretical spectra generated from an *in silico* proteolytic digest of a protein sequence. All spectra of the same precursor mass, within a given tolerance, are scored in terms of similarity against the experimentally obtained spectrum. The 'winner' is usually the best scoring spectrum and is taken forward to represent that precursor identification. The construction process starts with the theoretical digest of the sequence database. This is configured by the user and should reflect the conditions under which the sample was prepared. The enzyme used for proteolysis is selected along with the number of incomplete cleavages allowed. This determines the rules used to create the peptide strings forming the basis for building the theoretical spectra. The user must also supply a (limited) number of possible post-translational modification, the amino acid specificity and whether they should be applied to all occurrences of the given amino acid. This will influence the mass of the peptide and the modeled fragmentation patterns. The second step decides the *m/z* of the peaks to include in the theoretical spectra. This can vary according to which type of fragments are expected be present, the type of mass spectrometer being simulated and its inherent fragmentation properties and is also determined by the user to some extent. Depending on the desired complexity of the final model one can, at this stage, also consider applying rules to describe the fragmentation properties of component amino acids in the peptide sequence (for instance water or ammonia losses, mobile protons etc). The intensities of the peaks are next built into the model and fall into three categories of complexity[73]:

1. UT spectra (uniform theoretical spectra): all peaks have the same intensity
2. FT spectra (fragment theoretical spectra): the peak height varies depending on fragment type
3. RT spectra (residue theoretical spectra) different intensities assigned based on (learned) statistics about the fragment type, its position in the spectrum, and other fragmentation biases[74,75].

The construction of theoretical spectra is performed for the peptides resulting from an *in silico* proteolytic digest of all proteins in the query database (according to user preference) and stored, usually indexed in some way to the peptide along with any additional information such as its neutral mass, the number of miss cleavages, and any given fixed or variable modifications. With most search engines, all candidate proteolytic peptides matching within a given mass tolerance of the precursor *m/z* are firstly selected. They are then compared to the corresponding theoretical spectra and a similarity score is recorded. There are many different types of scoring scheme published, with non-probabilistic approaches including counting matching peaks[76,77], spectral correlation functions where the goodness of fit of the acquired fragment ions is compared to the model[78–81], rank based scoring[82] and finally probabilistic scoring[83,84], where matching might also take into account the *m/z* and the intensity of the acquired fragments, potentially penalizing unexpected, high intensity peaks. The score is then related to the probability that the model used could have arisen by chance alone from the search database used. This overall process is summarized in figure 8.

Extension of sequence database search: PTM localization and open searches One of the most important, and possibly underestimated, shortcomings of the sequence database search approach are the limitations presented by PTMs. The addition of a single PTM to a residue means that all theoretical fragment ions after the affected residue are shifted by the mass of the PTM. If the PTM is not 'fixed' (i.e. does not affect all residues) then a second set of potential fragments (those containing the mass shift as well as those without the modification) is added to the list of possible fragments to match to the observed fragments for that peptide. This increase in search space is then applied to all instances of the given amino acid in all peptides. With multiple potential PTMs of a kind on each peptide, and with many tens of potential PTMs, it is clear to see that searching with many PTMs can quickly lead to very long search times. Secondly there might not be enough experimental evidence to actually localize the PTM to one particular residue within the peptide. This has led to the creation of several re-scoring methods which use the presence of specific fragment ions to localize the modification to one residue and provide a score[76,86–88] These solutions greatly



Rank	Peptide	Score
1	QVSVVDLTNTR	3.3
2	VVEELCTPEGK	2.1
3	DLLLQWCWENGK	1.9
4	ECDVVSNTIIAEK	1.7
5	GDAVFVIDALNR	1.6
	…	
n	SYLFCMEAEK	0.2

Peptide	Score	Expectation value
QVSVVDLTNTR	3.3	$3 \cdot 10^{-4}$

Figure (8): From fragment to protein. The path of spectral matching from ref.[85]. A comparison between one or more theoretical fragment ion spectrum and an observed tandem spectrum yields some kind of similarity score. Usually the peptide corresponding to the best-scoring match is assigned to the acquired spetrum. An expectation value may also be calculated to illustrate the probability of the given match arising by chance from the database used.

improve the situation but cannot improve the situation if the underlying spectra are of poor quality. Of greater importance than the localization of a modification within a peptide is the issue of unanticipated modifications. It has been estimated that up to one third of spectra remain unassigned because of unexpected modifications[89], i.e. those which are not added to the search space in order to prevent explosion in possible MS2 fragment series to be compared (described above). There are two main ways to mitigate this problem: The first solution is to perform a subsequent 'data dependent' search. Here a second round of searching takes place limiting the peptide or protein [sequence] space to those peptides (or proteins) that were identified in the first round[90,91], but greatly increasing the numbers of potential modifications. The underlying assumption of both approaches is that the unmodified forms of peptides or proteins found during the first round are also likely to be present in the sample in one or several modified forms, but at levels substoichiometric to their 'bare' counterparts. In the ModifiComb approach, a comparison is made between all high confidence assigned peptides and the remaining unmatched spectra from the same acquisition file. The delta change in mass (ΔM) between a high confidence peptide and the unassigned spectrum is noted and the theoretical fragment ions of the assigned peptide (with the addition of fragments corresponding to the given ΔM) are compared to the fragment ions of the unassigned spectrum. If the number of matched fragments is above a given threshold, the previously unassigned peptide is considered matched and the ΔM between its precursor and that of the assigned peptide corresponds to a modification. Since fragments (plus those corresponding to ΔM) were included in the theoretical fragment ions, a possible location of the modification might also be obtainable. In the Mascot Error Tolerant search approach, a new search database is created containing only proteins identified with high confidence from the first round, which is used to search the unassigned spectra with a long list of potential modifications. Because only a limited number of peptide or proteins are searched the additional search space brought about by the increased modifications is still manageable. The second solution is described by the 'open search' approach. Here, the user performs just one search, setting a very large precursor tolerance i.e. ± 500 Da, but a small fragment ion tolerance. Whilst this does not reduce the search space for the candidate peptides, the second step (where theoretical fragment ions are compared to those observed) can be much quicker. The main advantage of this approach is that it does allow peptides to be identified which would have been otherwise been missed with the tighter precursor tolerance search. Recently a new method to substantially increase the speed of open searching has been introduced[92](by linking the individual fragment ions to their parent spectra through a high-performance index) which promises to ameliorate the PTM search space problem.

The advantage of sequence database searching is its relative simplicity: assuming you have a source for protein sequence information it will be possible to match a high percentage of the acquired spectra with a good level of confidence. However, if your species of interest is not well annotated or you are interested in a large number of potential PTMs you may need to look to other methods. The sequence database is very often the basis for the next two approaches which, on the whole, rely on previous peptide to spectrum matches to validate the experimentally acquired spectra which they use.

1.5.1.3 Spectrum matching The second method of obtaining PSMs from experimentally acquired spectra is via a spectral library. The principle idea here is that the fragmentation pattern of a peptide acquired under a given condition should be exactly reproduced when the peptide is fragmented at any given point in the future assuming it is acquired under the same conditions. It should then be possible to match the newly-acquired spectrum to this peptide fingerprint, thereby quickly revealing the peptide sequence. In reality, this fingerprint is slightly different between runs and it is therefore necessary to use a consensus spectrum built out of many high-quality, manually validated experimental spectra of the same peptide and modification status. The consensus spectra are stored in a library linked to the parent mass. The libraries are often built from dedicated peptide-finding shotgun experiments for a specific biological system or tissue type or, more recently, a vast set of synthetic peptides[93]. It may also be desirable to build a spectral library from repositories of raw spectra; attempts to build a centralized spectral library such as those at NIST[94,95] through PeptideAtlas[96] or PRIDE[97,98] have been running since 2006. It is important to consider that libraries should ideally only be built from data acquired under the same or very similar conditions to those under which the experimental spectra are acquired. There are several different search engines compatible with spectral libraries. Examples include spectraST[99], X!hunter[94] and Bibliospec[95]. Mascot version 2.6 also offers a spectral matching facility in addition to its standard database searching capability. Owing to the relatively small search space, spectral matching should be orders of magnitude faster than database or *de novo* searching methods and because all the features of the consensus spectrum are utilized it should also be more precise.

The scoring methods generally work by using a dot product to compare the sum intensities within a given *m/z* bin, hence there is more weight placed upon similar intensities rather than differences between theoretical and acquired *m/z*. This is the main difference from the database searching methods. Experimental data are often pre-processed prior to searching to remove noise based on arbitrary thresholds or the top*N* (number of) peaks. Significance is given to the matches in ways similar to database searching.

1.5.1.4 *De novo* and machine learning approaches It is possible to obtain PSM predictions without any existing knowledgebase. This is third type of PSM determination that will be discussed. *De novo* sequencing is method that relies solely on the information contained within the acquired MS/MS spectrum to determine the sequence of the peptide. This technique potentially allows to identify spectra that might be absent from a search database or equally enables discovery of a post-translational modification that might not anticipated *a priori*. It works by building a ‘ladder’ of mass differences between pairs of peaks in the MS/MS spectrum which, within tolerance, specify the mass of an amino acid. Such a ladder is built for at least the N- and C-terminal fragment series. The simplest form of *de novo* sequencing, known as the 'naive approach', builds the ladders by sequentially breaking down (and subsequently scoring) the residues from a list of candidate peptide sequences which match the precursor within a given mass tolerance. To keep the list of candidate peptides short enough to traverse in polynomial time, it is either limited to shorter peptides (of lower mass) or those matching within a very high mass accuracy. This approach is implemented in a commercial software PEAKS[100] where a limit of 10,000 subset peptides is set. The spectra are then scored based on the peak abundance, mass error and fragment complementarity. The second and most commonly-implemented type of *de novo* sequencing is the 'spectrum graph approach'. Here each spectrum is transformed into a directed acyclic graph through which the optimal path (ladder) is found, usually via dynamic programming algorithms, to sequence the peptide. Each peak in the spectrum is assigned to a vertex and the vertices are connected by edges if they differ by mass of an amino acid. The edges may be weighted according to peak intensity or intensity rank. Examples of this type of algorithm include PepNovo[101] and Lutefisk[102]. They use a variety of different scoring methods including model-based probability and simple ion abundance and mass tolerances. The main disadvantage of this approach is the requirement of good quality MS/MS spectra. A poor spectrum can result in a completely missed peptide match if it contains missing peaks or is very noisy. Therefore, mass spectra originating from higher accuracy mass spectrometers lead to better *de novo* predictions, but the matching is still complicated by the fact that the fragmentation of precursor ions is not always complete and may depend on the precursor abundance or the energy used for fragmentation. CID might suffer from missing ions in the lower and upper *m/z* region of the spectrum and a study by Chi and colleagues argues that HCD fragmentation offers some optimism in this regard since HCD spectra do not have this limitation, comprise high accuracy fragment ions in nearly complete ion series and also yield abundant internal and immonium ions[103]. They also state that it is possible to use combine data from ETD/HCD fragmentation of the same precursor, leading to improved results and published a tool, pNovo+, to perform this. However, the disadvantage desribed above remains and, coupled with the fact that *de novo* sequencing never links directly to a protein sequence, spawned the development of combined peptide matching / *de novo* methods. Here, a short tag is generated by the *de novo* approach and then used for searching within a database and is exemplified by the GutenTag[74] or UniNovo[104] approaches. In their Meta-SPS approach Guthals *et al.*[105] claim to have circumvented this issue by combining CID / HCD and ETD fragmentation of long, overlapping peptides obtained through multiple enzymatic digests.

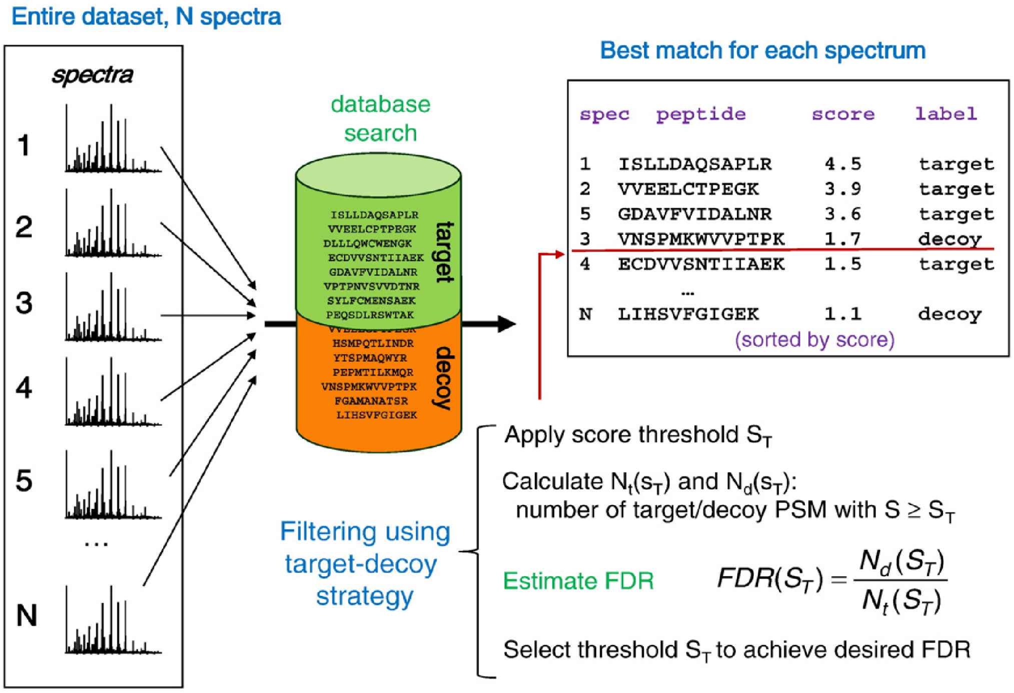
The other method of searching (to be mentioned only briefly here) is the application of machine learning methods to predict and or score spectrum to peptide matches often without the need of a theoretical sequence database. MS2PiP[106,107] is a tool to predict the intensities of the most important fragments based on a random forest machine learning approach using a large set validated peptide to spectrum matches.

It forms the basis for a search-engine-like tool which compares the intensity information in the predicted spectra with those recorded in the experimental spectra to sensitively identify true PSMs[108]. Other similar methods are also starting to appear in the literature[109,110]and promise to add a further dimension to the techniques available to match experimental spectra to theoretical peptide sequences. It is of course possible to combine different aspects of these techniques.

On the whole the sequence database search currently offers the most practicable solution for high throughput proteomics workflows. As a result, it has seen the largest number of applications developed. Its main drawback is the limitation of only finding peptides and PTMs which are present in the search database and defined in the search parameters. The reduced search space approach can still be used for most applications; projects requiring detailed analysis of specific PTMs or the emerging field of proteo-genomic analyses should also include other approaches such as *de novo* sequencing and certainly machine learning approaches.

1.5.2 Estimating peptide false discovery rate

Gauging the level of confidence one can place in the PSM matches and filtering the data accordingly is an essential part of the post-spectrum assignment process, which also enables one to compare result data originating from different labs and data which has been processed using different post acquisition methods. The concept was first introduction by Benjamini and Hochberg[111] and was adapted for proteomics with the publication of the target-decoy approach[112]. Here, the first step is to create a set of known false-positive PSMs which mirrors as accurately as possible the properties of the target database, but which does not exist in the target space, effectively the null model. The score distribution of these generated ’decoy peptides’ will follow that of the false positives. There are a number of ways to generate the decoy sequences, the commonest being to simply reverse the target protein sequences. However, some find this unsatisfactory since the mass distribution of the decoy sequences does not exactly mirror that of the target sequences because of the shift in location of enzyme cleavage sites. To counteract this, some argue that one should reverse or randomize the sequence within the proteolytic peptides, keeping the C-terminal residue the same so that the same peptide distribution is maintained. However, several studies have shown this does not lead to significantly different results[113,114]. There are two strategies for performing the searches: the first and most popular is to concatenate the decoys to the target search database and set up a ’competitive’ search between targets and decoys and the second is to perform two separate database searches (one against targets and one against decoys). Both methods have their drawbacks and some studies have reported both methods to be too conservative in their false discovery rate (FDR) estimation[113,115,116]. These inaccuracies can be countered by correcting for the competition effect[117]and more recently by other methods such as averaging the FDR after using different decoy databases[118] . Other estimation methods are available but the FDR approach, owing mainly to its great simplicity has been the most


Entire dataset, N spectra
spectra
1
2
3
4
5
...
N
database search
ISLLDAQSAPLR
VVEELCPTPEGK
DLLLQWCWENGK
ECDVVSNTIIAEK
GDAVFVIDALNR
VPTPNVSVVDTNR
SYLFCMENSAEK
PEQSDLRSWTAK
target
HSMPQTLINDR
YTSPMAQWYR
PEPMTILKMQR
VNSPMKWVVPTPK
FGAMANATSR
LIHSVFGIGEK
decoy
Best match for each spectrum
(sorted by score)
Filtering using target-decoy strategy
Apply score threshold S_T
Calculate $N_t(s_T)$ and $N_d(s_T)$:
number of target/decoy PSM with $S \geq S_T$
Estimate FDR
Select threshold S_T to achieve desired FDR

spec	peptide	score	label
1	ISLLDAQSAPLR	4.5	target
2	VVEELCTPEGK	3.9	target
5	GDAVFVIDALNR	3.6	target
3	VNSPMKWVVPTPK	1.7	decoy
4	ECDVVSNTIIAEK	1.5	target
	...		
N	LIHSVFGIGEK	1.1	decoy

$$FDR(S_T) = \frac{N_d(S_T)}{N_t(S_T)}$$

Figure (9): FDR estimation using combined search strategy, where all acquired tandem MS spectra are searched against a database containing both targets and decoy protein sequences and thereby compete to be the top hit.The FDR is calculated by applying a score threshold and summing the total number of targets and decoys with scores above that threshold. The desired FDR can be used to select the corresponding score threshold. Taken from ref.[120] with permission.

widely adopted in proteomic mass spectrometry workflows. Once the global FDR is set, filtering data can take place to remove all hits below the score associated with a particular FDR , or q-value[115]. The method described here is particularly suited to search database strategies but can also be applied to spectral library searching[119]

The above method seeks to identify a global FDR relative to a given peptide within the whole data set. It does not give any confidence value to an individual peptide to spectrum match. The post-error probability also known as local FDR can be calculated in a similar way to the global FDR and the two terms are indeed related[115].

The speed with which a search is performed is critical to its acceptance by the community as well as its scalability to perform in high throughput pipelines. In sequence database searching, the speed is dependent on the search space and complexity of scoring, with the former being the most influential. The search space is determined by the number of comparisons between the acquired spectrum and all candidate spectra. It increases with the size of the search database, decreased matching precursor tolerance and with a greater number of post-translational modifications (why the number supplied by the user should be limited). Not only is the search speed decreased but the false positive and possibly also the false negative rate increases. It can be solved to a certain degree by simply increasing computational power but also by building 'smarter' algorithms. Having accurately identified peptides with the required level of confidence, the next part of the puzzle attempts to take these protein-parts and re-create the original proteins present in the sample.

1.5.2.1 Post-search processing Following on from the steps outlined above, a list of PSMs is generated, each with an associated q-value and posterior error probability (PEP). At this point it is still possible to extract more information from these data, for example by combining the results of several search engines[121–123]or by feeding the results into a machine learning apparatus. Percolator[124], a semi-supervised learning approach, attempts to investigate the features of the high-confidence target peptides within the given run. These are then weighted for their influence on score. The features and weights can be applied to the PSMs discarded due to the given FDR cut off and re-scored to 'rescue' missed PSMs. It was shown that up to 17% more spectra can be correctly assigned using this method. Combination of the results of different search databases can be achieved using online tools such as PeptideShaker[121], IPeak[125], SearchGUI[126] or Ursgal[127] (see also section 1.5.6).

1.5.3 Protein inference

In the context of a typical peptide-centric MS workflow the PSMs gained above correspond to stretches of amino acids resulting from the proteolytic cleavage of proteins which were present in the starting mixture prior to being separated by one or more subsequent methods. The result of this is a set of peptide sequences each with a score, a mass and in most cases very little additional information about their origin. The task is now to use them to reconstruct the original set of proteins present in the starting mixture as accurately as possible. This undertaking is far from simple. Firstly, the lack of experimentally derived information relating to the intact protein and the possibility that one peptide sequence may map onto one or several highly similar protein sequences often results in a large number of shared or degenerate peptides. Secondly not all proteins in the original sample will have been present at equimolar abundances meaning that the peptides of more abundant proteins are likely to dominate over, or mask, those of lower abundance[128], and a mass spectrometer running in DDA mode is more likely to select the more intense precursor signals for fragmentation (see 1.3.5.1). Thirdly, not all proteolytic peptides have an equal chance of being detected by the mass spectrometer. These differences in their intrinsic detectability depend both on ionization efficiency (the fraction of gas-phase ions generated from the total number of molecules) as well as other factors, such as ion transmission efficiency and detector response. Fourthly, the difference in length of different proteins means that the number of proteolytic peptides is not equal and it is possible that large parts of a protein's amino acid sequence may simply not be conducive to proteolytic digestion by the enzyme being used. The combination of these factors complicates the task of protein inference.

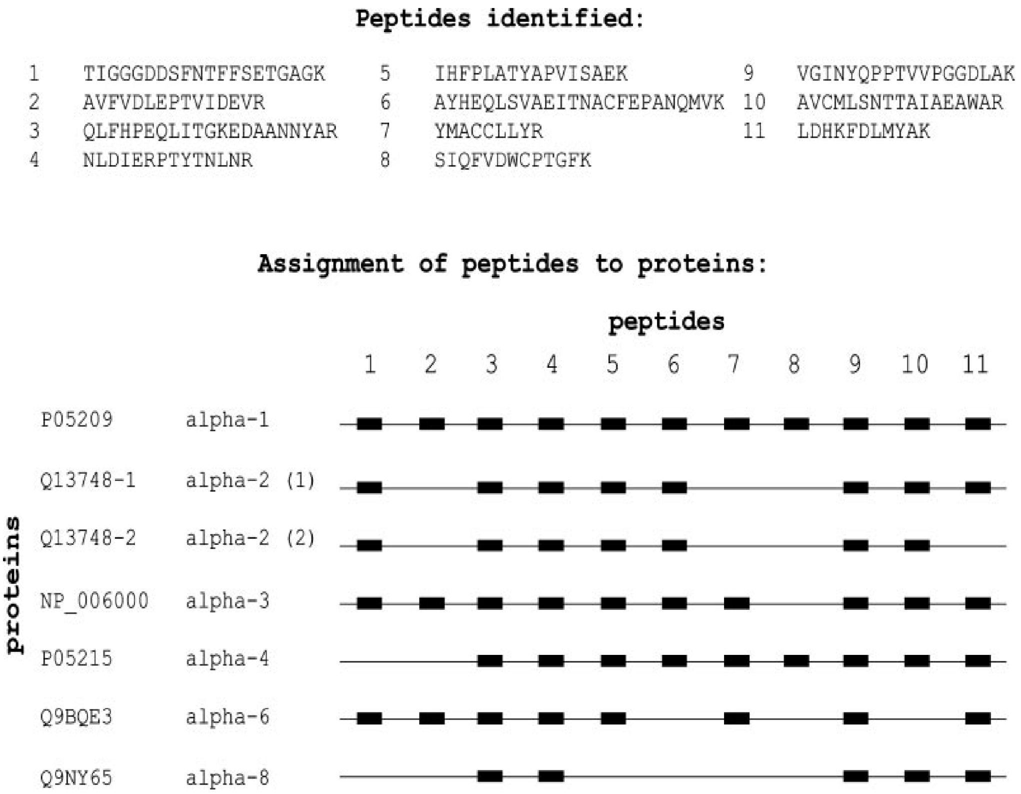
1.5.4 Determination of protein false discovery rates

The determination of protein false discovery rate and control of the uncertainty of protein assignments is much more of a challenge than at the level of PSM. This is mainly because we are dealing with assemblies of peptides, each with its own individual q-value and the fact that a protein makes the same contribution to the total protein distribution irrespective of the number of peptides it contains. This can be made worse by the fact that any errors coming from the PSM level will propagate up to the protein identification level. There is also the problem that true positive PSMs will map exclusively to the smaller subset of proteins present in the biological sample than the randomly matching decoy peptides. This can result in a protein FDR that is overestimated and much larger than the PSM-FDR and which deteriorates with increasing size of the dataset. The MAYU algorithm[129] attempted to correct for this over estimation by modeling the number of false positive protein identifications using a hypergeometric distribution. Its parameters are estimated from the number of protein database entries and the total number of target and decoy protein identifications made. The protein FDR is then estimated by dividing the number of expected false positive identifications (expectation value of the hypergeometric distribution) by the total number of target identifications. This goes some way to ameliorate the problem but has been improved on by Savitski *et al.* in the ’picked protein’ approach[130]. Here the overestimation of decoy proteins is prevented by pairing decoy and target sequences from the same source accession. In cases where both (target and decoy sequences) are identified in the dataset, only the best-scoring member of the pair is kept and the ’classic’ target-decoy search strategy (TDS) approach is applied.

1.5.5 Degenerate peptides and spectra

A peptide is described as degenerate if its amino acid sequence is shared with one or more other proteins. It can often arise because alternative splicing of exons from a single gene lead to alternative protein products sharing a large amount of homology across their sequences or equally as the result of different gene products from the same gene family or possibly, if it is a short peptide (less than seven amino acids) due to random matching. In any case, the peptide sequence is ambiguous and one cannot make any inference about the its origin. Figure 10 shows a protein with several isoforms and a large region of shared sequence. A degenerate spectrum can match to more than one PSM within the tolerance of the instrument it was measured on and may also come about because of residues of identical mass such as isoleucine and leucine or possibly asparagine / aspartate or glutamine / glutamic acid on a low-resolution instrument. It has been reported that these confounding peptides can be ignored for protein inference[131], but to improve accuracy this degeneracy must be taken into when performing protein inference, usually increasing the number of potential proteins a peptide could have originated from. An overview of the protein inference problem was first given by Nesvizhskii and Aebersold[132] in 2005 who concomitantly put forward a standard nomenclature to exhaustively describe all different peptide grouping scenarios (see Fig. 11) which, combined with Occam’s razor approach (also advocated by the same authors in ref.[128]) can provide a minimal list of explanatory protein identifiers for all peptides observed in the experiment. This list is at the lowest possible level of complexity (parsimonious) and presents the user with a conclusive set of proteins that were present in the sample and yields a consistent measure of the number of proteins identified in the experiment. However, this minimal list with a strict implementation of Occam's razor represents a limited view on the data: a researcher interested in a specific protein which is only present as a subset protein identification will miss their protein of interest. Therefore, the authors suggested listing all inconclusively identified proteins in addition to the minimal list. The converse to the minimal list above is referred to as the maximal explanatory set of proteins[73] or the optimistic model[133] and refers to the list of all proteins explainable by the observed set of peptide sequences. It is based on a set of theoretical peptides derived from the search database and the optionally also parameters used in the search (for instance, the number of missed cleavages or selected modifications). This set of proteins represents all possible alternative hypotheses, rejecting no possibility and ultimately leaves the decision as to which protein(s) were present in the original sample to the downstream user. A scan of the literature surrounding protein inference since the review by Nesvizhskii reveals that many algorithms offering different ways to derive explanatory protein lists have been published. Most are still heuristic but come equipped with scores to assess the confidence of the inference made and many give an estimation of the FDR (or the expected proportion of 'wrongly inferred' proteins) amongst the significant hits. In a recent publication, The and co-workers state that these methods differ mainly in how they deal with the degenerate peptides[134] and are divisible into three groups: inclusion, exclusion and parsimony. An inclusion method infers the presence of any protein which links to an identified peptide in a similar way to the maximal explanatory set method; an exclusion method which entirely removes any shared peptides to base protein inference entirely on the peptides unique to one protein; and parsimony as described in the original suggestion by Nesvizhskii and Aebersold. They tested representative algorithms from each of the three ways to treat shared peptides, using five different scoring methods to infer proteins acquired from three different protein mixtures. Finally, they conclude that inference procedures excluding shared peptides provide more accurate estimates of errors compared to methods that include information from shared peptides, while still giving a reasonable performance in terms of the number of identified proteins. However, they did not include large protein groups in their assessment. The discussion regarding protein inference is therefore far from over. Serang states that the field of computational proteomics should move away from heuristics and approach problems such as protein inference formally[131]. He says that using more complex models in mass spectrometry will introduce a greater computational burden but that should not dissuade us from modeling the process as accurately as possible: using, for example, prior information about the proteins present in the sample or regarding peptide detectability, a sentiment echoed by Li *et al.*[135]who support potentially exploiting peptide detectability predictions and better estimates of protein/peptide quantity using a Baysian approach[136].

The stochastic nature of DDA MS experiments leads to a probable overlap in protein identification of 70-80% from one experiment to the next, assuming that exper-



Peptides identified:

1	TIGGGDDSFNTFFSETGAGK	5	IHFPLATYAPVISAEK	9	VGINYQPPTVVPGGDLAK
2	AVFVDLEPTVIDEVR	6	AYHEQLSVAEITNACFEPANQMVK	10	AVCMLSNTTAIAEAWAR
3	QLFHPEQLITGKEDAANNYAR	7	YMACCLLYR	11	LDHKFDLMYAK
4	NLDIERPTYTNLNR	8	SIQFVDWCPTGFK		

Assignment of peptides to proteins:

		peptides										
proteins		1	2	3	4	5	6	7	8	9	10	11
P05209	alpha-1	■	■	■	■	■	■	■	■	■	■	■
Q13748-1	alpha-2 (1)	■		■	■	■	■			■	■	■
Q13748-2	alpha-2 (2)	■		■	■	■	■			■	■	
NP_006000	alpha-3	■	■	■	■	■	■	■		■	■	■
P05215	alpha-4			■	■	■	■	■	■	■	■	■
Q9BQE3	alpha-6	■	■	■	■	■		■		■		■
Q9NY65	alpha-8			■	■					■	■	■

Figure (10): Illustration of a protein family: a set of highly degenerate (overlapping) peptides, none of which are unique to any single protein, means that one can at best record that the protein family is identified in the experiment or at worst that no unambiguous identification is possible. Taken from ref.[132] with permission.

imenters follow a standard operating procedure (SOP) and use instrumentation at peak performance[137]. This may still result in proteins being 'missed' in one experiment, but could potentially be solved by the approaches described by Li and Serang or possibly by re-inferring proteins based on peptides from both / all experiments. This is the current approach taken by software like MaxQuant[138]. It might also be sufficient to just group peptides by their annotated / associated genes. Many algorithms try to answer the unanswerable: without additional information gained from further experimental validation, it is impossible to accurately reconstruct the set of proteins that was present in the original mix. Depending on what the desired outcome of an experiment is, the level of protein grouping should be adapted and, in specific cases, two types of explanatory set reported. The boundaries for the discussion of protein grouping would again be changed if the constituent peptides carried more information. This information could, for example, take the form of a quantification ratio (as was already suggested by Nesvizhskii in 2005). For now, the burden is on the user to decide, within the context of the biological question, which level of protein inference should be sought and how false positives should be dealt with.

1.5.6 Data workflows.

The steps outlined above cover only part of the entire flow of data from its acquisition on the mass spectrometer through several conversion steps, filtering of peptides, protein inference, to peptide and protein quantification, with relevant quality control

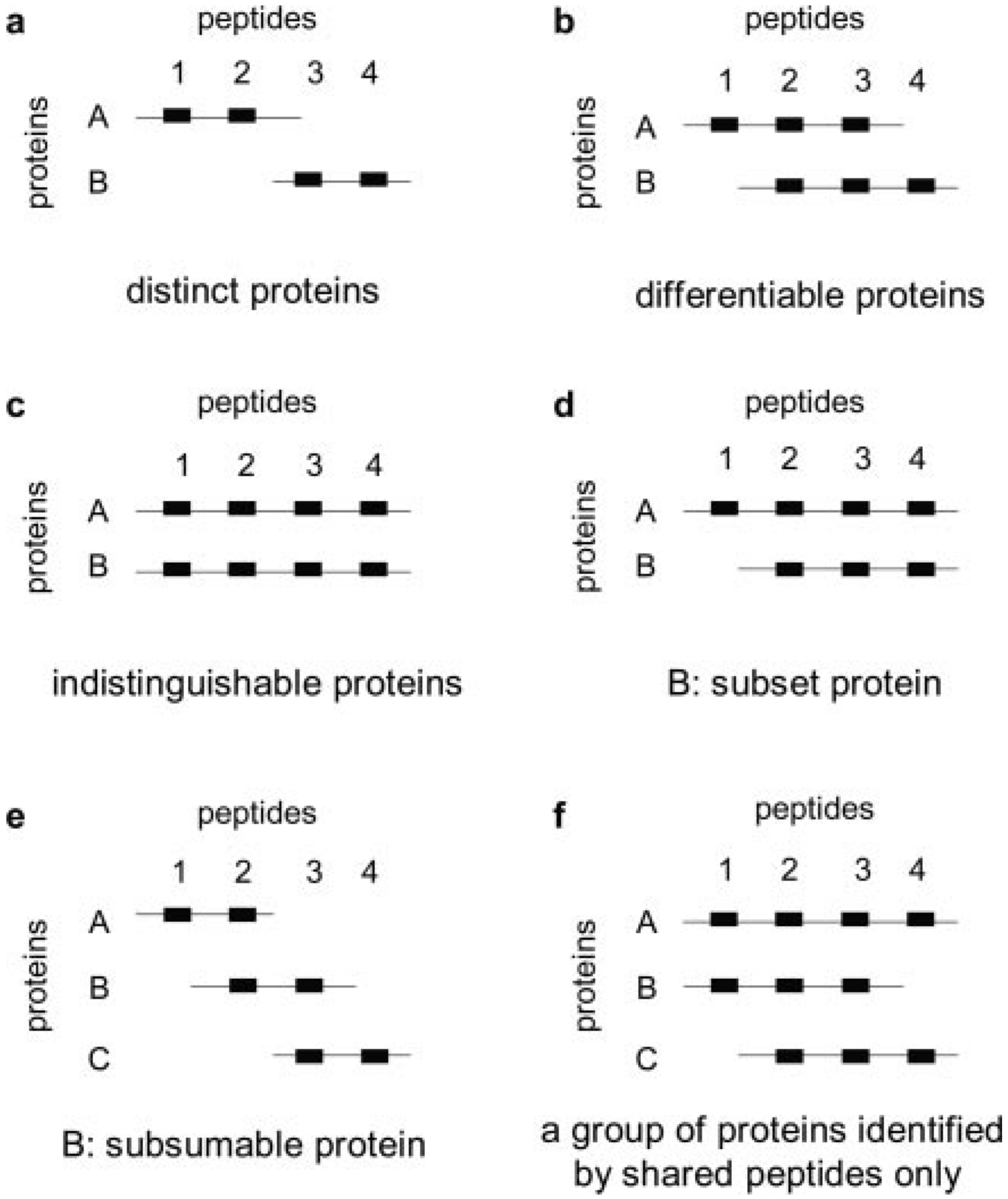


Figure (11): A suggestion of six definitions for peptide grouping scenarios taken from ref.[132] with permission.

(QC) of data at each step. The destination of the journey of the data is to create an output which is simple for biologists and experimenters to interpret and which provides a simple interface to any downstream tools and which has the necessary links back to the original conditions that were present prior to the acquisition of data. There are currently a number of tools and workflows which are available for processing Mass Spectrometry data. A list of example software covering a large part of the end-to-end data workflow is given below:

1. MaxQuant[138] is a self-contained, closed-source, freely available software. It was originally published only for Windows but was recently also made available for a Linux operating system[139]. It has its own in-built search engine (Perseus) which was released as part of comprehensive workflow-based data analysis platform[140] and was later also given command line functionality. It has an easy to use with a graphical user interface. It offers in-built data visualizations for QC and the main output(s) are text files.

2. TPP[141] is available for Linux, Windows and macOS. It first of all converts different vendor formats into a HUPO standard extensible markup language (XML) format (mzML - http://www.psidev.info/mzML), which is then fed to the next part of the workflow where an array of different (external) search engines or spectral libraries map the experimental spectra to peptides. It then performs peptide quantification followed by protein assignments. The outputs are visualized via a web browser.

3. OpenMS[142] is a suite of software tools, written in C++ with bindings to Python. It runs under Windows, Linux and macOS. Its workflows are constructed in a similar way to KNIME[143] and there is currently an initiative to combine these tools.

4. Proteome Discoverer[144] is a commercial software provided by Thermo Fisher Scientific. It is similar to pipeline pilot and KNIME in that it operates on a node-system where users can build up workflows by slotting together individual nodes which perform manipulations on the data before passing it to the next node. It is closed source and commercial, requiring a license to design workflows but several user licenses are free.

5. Skyline[145] describes a whole suite of software running on a Windows operating system primarily for building quantitative methods using SRM / MRM, PRM, DIA and DDA with MS1 and subsequent analysis of the resulting mass spectrometer data.

6. Galaxy P[146] is not a local but rather a global workflow tool with its roots in genomic informatics. Galaxy provides a user-friendly, web-based, scalable platform where disparate software tools can be integrated into useful workflows. It is accessible through Jetstream, a cloud-based scientific computing infrastructure.

7. Ursgal[127] is a tool written in Python running on any platform operating system. It performs searches for a wide variety of search engines and combines the results prior to running Percolator[124] on the unified results. It supports other workflows within the computational mass spectrometry space such as the creation of search databases with decoy peptides using Ursgal[127].

8. PeptideShaker[121] is an online portal / freely-downloadable Java-based tool for the interpretation of proteomics identification results. It combines the results of multiple search engines in an attempt to extract the greatest amount of understanding from the interpretation of the acquired data. It does not perform any extraction of raw data, but is able to re-calculate PTM localization scores, carry out protein re-inference, perform gene ontology (GO) enrichment analyses and create QC plots.

1.6 Mass spectrometry data repositories

The ultimate destination of the results of a mass spectrometry experiment and the biological interpretation drawn from them will always depend on the lab where they were acquired. Fortunately, it is now requisite to deposit the data in large repositories such as PRIDE[98,147] when publishing findings of an experiment whose readout is mass spectrometry-based. This not only enables other researchers to access the raw data and draw their own conclusions from it but it has also enabled some major mass spectrometry projects, such as the draft of the Human Proteome[148] and inception of the ProteomicsDB[149] to be possible. Other repository members of the ProteomeXchange[150,151] consortium, whose main aim is to globally co-ordinate and standardize proteomic data submission and dissemination, include Peptide Atlas[96] (with a sub-repository PASSEL[152] focusing on SRM data), MassIVE (http://massive.ucsd.edu), JPOSTrepo[153], and Iprox[154]. However it is outside the scope of these repositories and databases to capture the final conclusions of the individual biological experiments and any decisions made based on them in a controlled way, for this experimenters have to read and interpret the accompanying literature themselves.

2 Objective and Aim

The main aim of the project described in this thesis was to create an open-source, freely available software tool that was capable of performing the required steps to extract, process and visualize the raw output of the Orbitrap family of (hybrid) mass spectrometers in order to give researchers the possibility to perform isobaric and MS1-based quantification on a variety of different starting materials in an accurate and precise manner. This should allow very accurate fold change determination even for large ratios. Support for experiments running in non-quantified or label free modes should also be provided, though quantification in a label free mode is outside the immediate scope of the tool.

The output of the software should be as accurate and precise as possible, being robust towards outliers and should include a level of confidence in protein and peptide assignments as well as for the protein quantification. It should be possible for the user to select the relevant way to group peptides according his or her requirements and the nature of the project; and grouping read outs from more than one mass spectrometry acquisition should not pose any challenge. Detailed information should be available about how the data in the outputs was derived but this should not overshadow the results themselves.

The tool should be designed to be fully open-source and freely available to the community and should incorporate expertise gained in processing of mass spectrometry data. The implementation should be designed in such a way as to combine the raw data with the processed data in single, self-contained file from which text outputs may be generated. The outputs should be as generic and uniform as possible to allow downstream processes to work, regardless of quantification type. The application must run in a command line manner so that it can be built into other pipelines or workflows with ease, and should run equally well on an experimenter's desktop as on a large server. The installation should be very simple for all levels of user and the software should be easy to download and it should work out-of-the-box with little or no further configuration required. It should be easily extendable to support different quantification modes.

At the beginning of this PhD project, no single tool was able to deliver, via a command line, all these objectives and include with it isobaric quantification, coupled with correction for ratio compression; accurate pulsed SILAC quantification and provide a single standardized output for direct manual interpretation, with the possibility to programatically interact with the raw and interpreted mass spectrometry data.

The name of the tool developed by the author to fulfill these objectives is called isobarQuant.

Part II

Creation of a stand-alone tool (isobarQuant) for processing and quantification of isobaric-tagged peptide data

1 Introduction & background

Isobaric tagging of peptides occurs post-experimentally, which allows investigators to perform experiments in as near to physiological settings as possible and then compare up to eleven different conditions in one mass spectrometry experiment. Because all conditions are analyzed together, the resultant read-out does not suffer from the disadvantages inherently associated with separate (detached in space and time) acquisition modes. The effectiveness and usefulness of this tagging strategy had already been widely evaluated and was at the heart of several high profile whole-proteome studies[155,156], biomarker research[157,158] and in PTM quantification[159,160], but at the time of starting this project there were very few freely-available software tools available which included dedicated support for isobaric tagging workflows. The intention of isobarQuant was to capture the experience and software requirements for working with isobaric tags and to publish this as an open-source, easily-downloadable software for the community. This was achieved in 2015[161].

The Python programming language was chosen to provide the basis for the isobarQuant package since it was already familiar and allows code development in fewer steps compared to Java or C++. The interpreted, object-oriented language comes with a large, comprehensive standard library that has automatic memory management. It is quick to prototype in and because it makes use of many libraries written in C, the decrease in performance that might be expected with an interpreted language does not present any issues. The workflow starts with the Thermo .raw file and terminates with a text-based output of protein fold changes. For ease of development isobarQuant was developed primarily as a command-line tool which, owing to the operating system limitations of the Thermo Fisher Scientific .raw files at the time of development, would primarily run on a Windows platform and should work equally well on a desktop PC, laptop or Windows server. At the time of writing it is being used by at least five research groups across the world and the GitHub repository (https://github.com/protcode/isob) gets more than ten unique visitors a week.

1.1 .hdf5 file format and PyTables

One of the most important decisions was trying to link raw data with later identification and quantification information in a single file, the size of which should, ideally not be greater than that of the original file. For this reason, the .hdf5 file format was chosen as it is high density, can be easily visualized and importantly, specific segments of the data may be accessed directly, via indexes, from disk as though being stored in physical memory yet residing on the hard drive of a computer or server. This leads to substantial improvements in performance and allows huge numbers of rows (up to 2 x 10^{63}) to be stored in one table.

.hdf5 files are organized in a hierarchical way, with two primary structures: groups and data sets. As the name suggests, the .hdf5 group is a grouping structure containing instances of zero or more groups or data sets, together with supporting metadata. A group is further divided into two parts: a group header, which contains a group name and a list of group attributes and a group symbol table, which is a list of the .hdf5 objects that belong to the group. The organization of groups can be described to be similar to a UNIX file system, with '/' sitting at the root. A dataset also consists of two parts: a multidimensional array of data elements and a supporting header containing metadata.The header contains information that is needed to interpret the array portion of the data set and includes the name of the object, its dimensionality, its datatype, information about how the data itself is stored on disk and further information used by the library to speed up access to the data set or maintain the file's integrity.

There are four essential classes of information in any data set header:

1. Name: a sequence of alphanumeric ASCII characters.

2. Datatype: consists of two categories atomic (integer, float and, string) and compound (made up of atomic data types)

3. Dataspace: describes the dimensionality of the data set and may be fixed or unlimited (i.e. extendable)

4. Storage layout: default is contiguous, meaning that data is stored in the same linear way that it is organized in memory or chunked

It is very easy to extend and update an .hdf5 file with almost any kind of data. In this regard it is essentially a mini database. At the time of development of isboarQuant, the main application programming interface (API) for Python to .hdf5 was PyTables[162]. PyTables not only provides the interface to the .hdf5 file type but also includes support for several python libraries required in the manipulation of large or vectorized datasets (e.g. numpy and numexpr), data compression (using the Zlib, LZO, bzip2 and Blosc compression libraries) out of the box. PyTables creates an object tree entity in memory which, upon .hdf5 file creation, represents the .hdf5 structure on disk. This is updated dynamically while the actual data is saved to disk. PyTables has three main classes: the node, group and leaf classes, with group and leaf classes being descendants of the node class. The group class roughly corresponds to the .hdf5 group described above with the leaf class encapsulating the properties of the data set (above). Leaves may not contain further groups or leaves and provide the

```
result = [row[’col2’] for row in table if ( ((row[’col4’] >= lim1 and row[’col4’] < lim2) or

((row[’col2’] > lim3 and row[’col2’] < lim4]))

and ((row[’col1’]+3.1*row[’col2’]+row[’col3’]*row[’col4’]) > lim5) )]
```

```
result = [row[’col2’] for row in table.where( ”’(((col4 >= lim1) & (col4 < lim2)) | ((col2
> lim3) & (col2 < lim4)) & ((col1+3.1*col2+col3*col4) > lim5))”’)]
```

Figure (1): Two examples of an .hdf5 file query using PyTables. Both queries work from on-disk data and only require the line being iterarted over to be loaded into computational memory. The ’table.where’ construct in the second example uses the in-kernel search facility and will be faster than the first. Columns in the .hdf5 file table are signified by ’col’ and limiting conditions are illustrated by ’lim’.

base class for the Table, Array, CArray, EArray, VLArray and UnImplemented PyTables classes.

1.1.1 Data selection using PyTables

Of critical importance to the success of isobarQuant is the ability to quickly and efficiently access a specific record (from disk) without the need for loading the entire contents of its parent table or dataset into memory. Parsing large, unindexed text or XML files are generally not well-suited to this task. There is a diverse array of possibilities for searching using PyTables and a couple which are pertinent to isobarQuant are outlined here. The simplest type of query is termed a Python selection and is preformed on a PyTables row object. Here the table is accessed row by row and the elements are selected according to the criteria given after the ’if’. It is still beneficial because only each row is loaded into computational memory at a time but it is likely to be relatively slow. It may be sped up by using the ’in-kernel’ search facility of PyTables (which uses the PyTables kernel module: an in-built C-module combined with the numexpr package), which enables the selection to take place before the iterator is returned. For selections that return relatively few records compared to the total number of rows the potential savings are large. This yields an up to ten fold improvement and might also be up to five times quicker than similar queries performed on a (non-indexed) postgreSQL database[163]. However, these improvements will not be as great for datasets which do not fit into dynamic memory at once, but still perform over two times faster and are just as easy to implement compared to the regular statement.

In a similar way to regular relational databases, indexing specific columns can also increase search speeds because it allows a binary search to be employed rather than a sequential search as described above. This is implemented in PyTables using a simple, single command and does not usual take too long to perform. PyTables offers different levels of index optimization according to the requirements. Once again, compared to a relational database, PyTables can be much faster since it is optimized for read-only or append only tables and puts a lower focus on updates and deletions. PyTables indexes also require much less disk space. One final feature of PyTables indexing worth mentioning here is use of ’sorted tables’, where one or more columns are (re)sorted and stored contiguously on disk. This can offer speed ups of up to one hundred fold when compared to an unsorted table with the same syntax as used in the second example above. Once again the time taken to create a sorted table will depend on the size of the data being stored and is aimed at tables which are largely read-only and will be accessed many times. The PyTables library is critical to the success and performance of any isobarQuant run. There are many factors which will influence the speed with which the file is generated, how fast it will be accessed and how large the final output files are, and some of these may be affected by external factors, such as disk type and read/write speed. Using PyTables is an excellent way of finding the optimal solution.

1.2 Data extraction

The Thermo Fisher Scientific Xcalibur tool offers an efficient method for experimenters to access the results of the MS runs that have just been acquired. It allows the user to see meta data relating to how the sample was acquired (what settings and parameters were used) next to the results of those settings and interact with the data as required. However, at the time of starting this PhD project it was not possible to access this information using generic software and moreover it was not possible, computationally, to compare the results of multiple runs. The data extraction and storage allowing quick access was a key design criterion for isobarQuant. It was desirable to record as many of the instrument settings and acquisition parameters as possible adjacent to the results they produced. This concept extends to Mascot search results and spectrum, peptide and protein quantification. Not all users will be interested in the entirety of this wealth of information; therefore only the most pertinent data should be later exported in text format for further analysis.

1.2.1 XIC trace extraction and reassignment of precursors masses

In order to get the best estimate of the precursor peptide mass (and not rely solely on the value recorded at acquisition time), it is desirable to select extracted ion chromatogram (XIC) traces for a range of stable isotopes of a precursor over their elution profile and match them to an expected value. This can give greater confidence in the precursor masses assigned. It is also beneficial to re-assign the precursor mass to the monoisotopic mass in cases where the ^{13}C peak was picked for fragmentation. XIC extraction is also required for the ‘Top 3’[50] approach to peptide quantification.

1.2.2 Deisotoping and deconvolution

The advent of HCD fragmentation and the increase in resolution and accuracy that accompanied the release of the LTQ-Orbitrap Mass Spectrometer required new ways of processing the tandem mass spectra to get the most out of the information contained in a spectrum. It had already been shown[164–166] that some level of pre-processing of spectra to remove noise peaks, additional isotope peaks and to spread out the ions across the entire spectrum space (via deconvolution) was beneficial for search engines, in particular for Mascot[167], which does not necessarily consider all peak intensities in a spectrum when calculating the peptide score. This would be a required functionality for the isobarQuant software.

1.2.3 Result file merging

An offline fractionation step prior to acquisition can reduce sample complexity and leads to improved proteome coverage by increasing dynamic range through a reduction in duty-cycle overload[168] and can also mitigate ratio compression[169]. However, merging all files prior to Mascot search can lead to very large, unmanageable Mascot result files. It was therefore a requirement of the software that isobarQuant be able to perform sample merging of two or more of the resultant .dat files in order to ensure that all peptides were recovered and that protein inference was correctly performed to accurately reconstruct the data of the original sample.

1.3 TMT-tagging of peptides and potential pitfalls

As described in the main introduction isobaric tags offer the possibility to label a high number of samples for acquisition in the same mass spectrometry experiment. There are several aspects which must be taken into account when performing TMT or iTRAQ quantification. Reporter ions should be extracted from MS/MS spectra and if necessary re-calibrated against the supplied masses. The level of potential reporter ion coalescence brought about by high numbers of ions in the Orbitrap should be estimated and accounted for when acquiring data[170]. Reporter ions with high levels of coalescence should not be used for quantification. The ability to correct the reporter ion intensities for mis-estimation resulting from naturally occurring isotopic distributions (from either manufacturer-supplied values or in-house determined experiments) should be included in simple way. It should also be possible to determine the level of peptide co-elution and correct for ratio compression[171]. A further requirement was the ability to carry out isobaric peptide quantification using the synchronous precursor selection (SPS)[35] method. It was essential to consider how to include all of the aspects mentioned here during the development of the isobarQuant software. It was also very important that the addition of new quantification methods be as simple as possible, and exclusion of one or more of the full complement of the available reporter ions not present any problem.

1.4 Protein inference & protein annotations

The difficulties associated with protein inference and problems facing experimenters performing bottom-up proteomics were discussed at length in the main introduction (1.5.3). This situation may be exacerbated when it comes to protein quantification: on one hand we wish to keep as much information relating to protein fold changes as possible to increase statistical power but on the other hand, peptides ambiguously matching to more than one protein (group) should be excluded from quantification as they might skew the signal with information that actually relates to a different protein (group). One potential way to circumvent this is to group proteins which are encoded by the same gene. The gene information is often provided as part of the entry in the search file used by Mascot. This gene information should be correctly recorded at the right place and then used in the determination of protein groups. The use of a gene name should also provide a stable identifier which can be used to compare across experiments, samples or larger cohorts. This also ensures that peptides mapping to different isoforms of the same gene are not excluded from quantification of that gene. The isoform-level information will not be discarded but remain available for inspection if required. This method of grouping can be switched off where necessary, returning the protein inference to the classic one protein-sequence procedure.The peptides used in the determination of protein sets and for protein inference must be recorded.

1.5 Protein quantification

There are several proposed ways to perform peptide-based protein quantification. With the primary focus of the first version of isobarQuant being on TMT quantification, the issue of missing values and the need to impute values was less important (see Part III for a detailed discussion of MS1-based quantification using isobarQuant and SILAC). It should, of course, be possible to exclude peptides from the protein fold change calculation if they fail due to defined filter criteria. The level of confidence in the protein fold change derived from the peptides should also be reported in order to gauge how well the value given describes reality.

1.6 FDR estimation at protein and peptide level

As mentioned in the main introduction, it is essential that both peptide and protein identifications are provided with an associated level of confidence. The best method for this is the implementation of a false discovery rate (see 1.5.2). A q-value can be calculated for each PSM, using the classical TDS approach[112] and for the protein identifications via the 'picked' protein approach[130] (see also 1.5.4).

2 Methods and implementation

Since spectrum to peptide identifications are made using the Mascot search engine, the workflow splits into two logical parts – the pre-Mascot and post-Mascot workflows.

2.1 Pre-Mascot workflow

The starting point for isobarQuant is the .raw file generated by the mass spectrometer. Currently isobarQuant is only set up to deal with files originating from the Orbi-

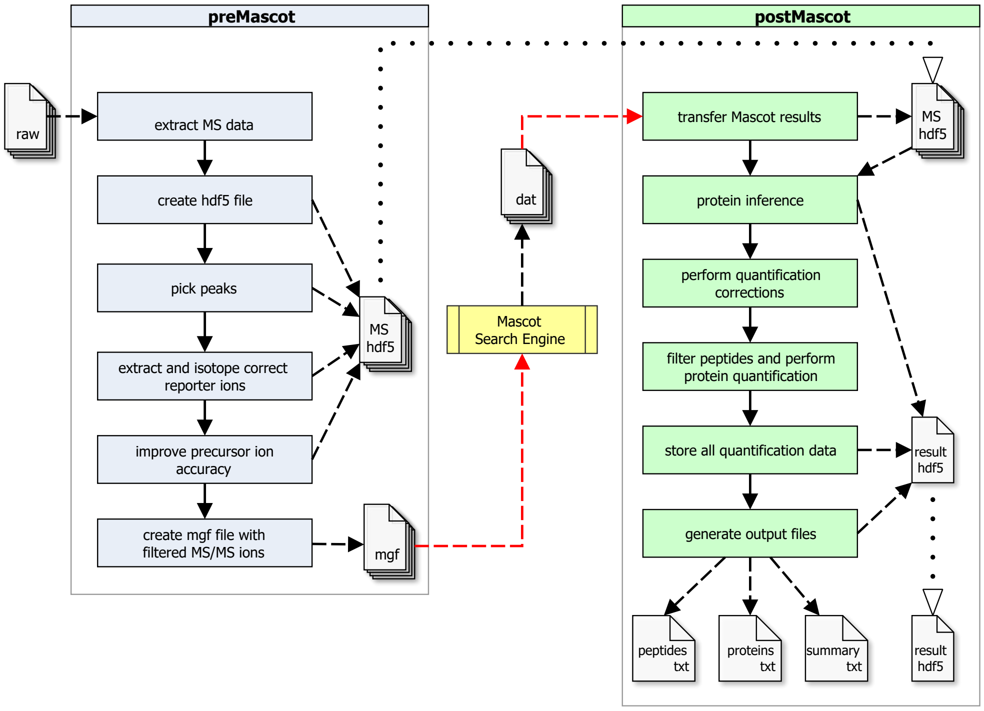

preMascot
postMascot
raw
extract MS data
create hdf5 file
pick peaks
extract and isotope correct reporter ions
improve precursor ion accuracy
create mgf file with filtered MS/MS ions
MS hdf5
mgf
dat
Mascot Search Engine
transfer Mascot results
MS hdf5
protein inference
perform quantification corrections
filter peptides and perform protein quantification
store all quantification data
result hdf5
generate output files
peptides txt
proteins txt
summary txt
result hdf5

Figure (2): Schematic showing the isobarQuant workflow starting with a newly acquired .raw file and ending in proteins and peptides with quantification information attached. The only manual step required is the submission of .mgf files to the Mascot search engine server. Taken from ref.[161] supplementary manual.

trap suite of Thermo Fisher Scientific instrumentation which includes Lumos Fusion, all types of Q-Exactive and the LTQ-Orbitrap series. isobarQuant interacts directly with the .raw file, without conversion to an intermediate format which allows the raw data to be stored adjacent to any data processed from it. It terminates with the generation of an .hdf5 file containing relevant data extracted and processed from the .raw file and an .mgf file suitable for Mascot searching. The .hdf5 file plus the result of the Mascot search (.dat file) are the basis for the second part of the workflow.

The pre-Mascot workflow of isobarQuant is started with a single command line parameter where the directory of the .raw files to process is provided. The regular expression facility of Python's pathlib module is used to find all .raw files located in a given directory. Each .raw file will be processed internalized in turn, the result being an .hdf5 file. Depending on the system setup, this could be lengthy for a large number of files. Because of this, a multi-thread option was developed to paralellize the work stream. See section 2.3.5.

2.1.1 Creation of .hdf5 file

A dynamic link library (DLL) written in Python is used as a wrapper to a C++ library provided by Thermo Fisher Scientific to access .raw files in a Windows environment. At the time of development this was the only API which was available. This API is called and relevant data from the .raw file is extracted as required. The first set of data is the acquisition / instrument parameters such as the order and type of scan events, the activation type, normalized collision energy used and the detector recording the

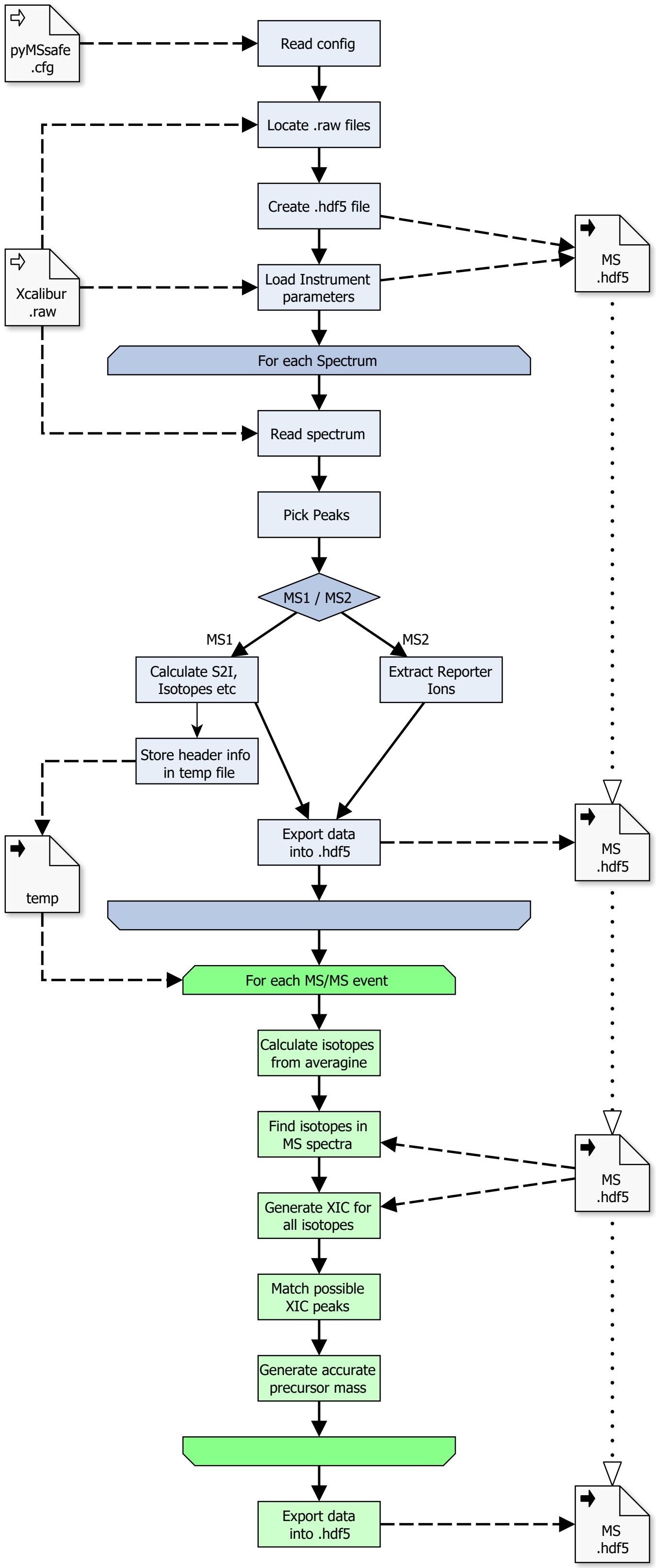


Figure (3): Overview of the PyMSsafe workflow. From top to bottom: the pipeline extracts data from the .raw file, starting with the acquisition parameters. Next, all spectra in the file are peak-picked and then processed according to their type (MS1 or MS2). The proportion of total intensity attributable to the precursor and from other peaks is calculated along with instrument noise threshold. This is stored along with the MS1 header data in a temporary file. Reporter ions corresponding to the isobaric labels are extracted, any potential coalescence is measured and these are stored in addition to the rest of the ions in a separate table. For each of MS/MS event, XICs are generated for isotopes of the precursor based on an averagine model and accurate masses is generated. Taken from ref.[161] supplementary manual signal are all written to the .hdf5 file and are stored in several tables according to type. Secondly all spectrum data is extracted and written to appropriate structures in the .hdf5 file. According to the information in the .raw file the, distinction is made between different types of scan events, for example the MS1 survey scan, the linked MS2 spectrum or multinotch spectrum: these associations are recorded. The type of spectrum also triggers different data processing steps depending on the acquisition mode and instrument used. These are described below.

2.1.2 MS2 smoothing and recording of ion intensities

For each MS2 (and, where applicable, MS3) spectrum, the data are smoothed by fitting a Gaussian model and the intensity of the top of each peak is recorded and summed for all peaks in the spectrum. The instrument-determined total ion current (TIC) for the spectrum is stored alongside the summed peak-top intensity values, which are later divided between the total signal derived from reporter ions and all remaining ions. More parametric data relating to how the spectrum was acquired, such as trap fill time, are recorded in the .hdf5 file as well as the number of MS2 scans from the corresponding MS1 survey scan.

2.1.3 Signal-to-interference and noise threshold extraction

An estimate of the purity of isolated peptide ions can be obtained by integrating the amount of signal coming from the precursor (and its isotopes) within the *m/z* range of the given isolation width and dividing it by the sum of all ion signals with the isolation window[172]. This signal to interference (S2I) was shown to be improved by incorporating the S2I value for the proceeding MS1 spectrum by extrapolating the S2I value as a time-weighted linear combination of both values[173]:

$$S2Im = (RTm - RTe)\frac{S2Il - S2Ie}{RTl - RTe} + S2Ie$$

The Xcalibur software (Thermo Fisher Scientific) applies a cut off to remove ions originating from chemical and electronic noise defined as all ions falling below 2.4 standard deviations of all detected signals over several sections of each spectrum[173]. This means the S2I estimation for these precursors will be inaccurate and should potentially be filtered away. This precursor intensity to threshold (P2T) is calculated by dividing the precursor abundance by the corresponding ion noise threshold and is extrapolated in the same way for each MS2 spectrum as for the S2I value[173]. The calculation of S2I and P2T is depicted in figure 4.

The calculation of S2I and P2T are particularly relevant when managing so-called ratio compression which is associated with isobaric, MS2-based quantification methods. Their use for other purposes remains unexplored (one could use it as a parameter in Percolator for example or as a trigger for 2nd (chimeric) MS/MS detection)

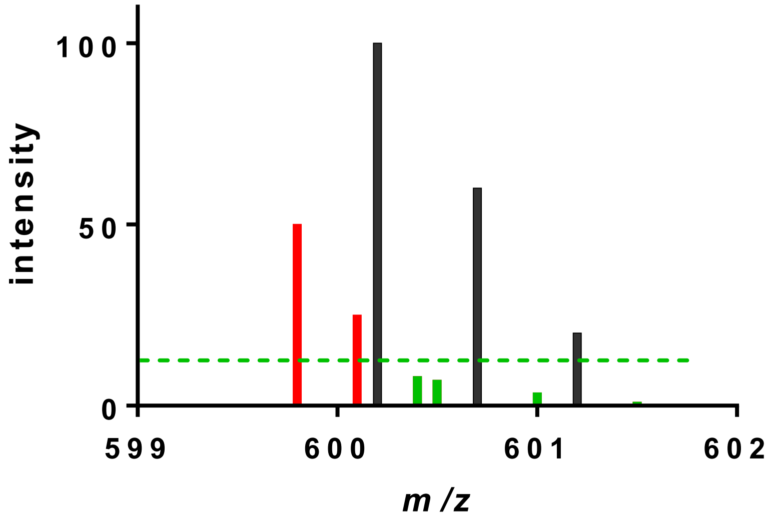


Figure (4): Assessment of S2I and P2T. The black bars depict the isotopic distribution of the precursor which was selected for fragmentation and whose reporter ions will be recorded and later used. Red (and green) bars show signal from the potentially co-eluting peptides. The dashed green line is the limit of detection applied by the instrument. All signals below this intensity are not recorded. A precursor intensity close to this cut off is likely to have a poor estimation of S2I, with more co-eluting peaks present than is accounted for. The reporter ions of such spectra should not be used for quantification.

2.1.4 Reporter ion extraction, correction and coalescence estimation

If isobarQuant is set to run in isobaric tag quantification mode (currently TMT or iTRAQ) each MS2 spectrum is queried for isotopes corresponding to the reporter ion masses given in the relevant configuration file. These are inspected for potential calibration offsets and assessed for potential coalescing ions as described in ref.[170]. Here ion signals are extracted from the MS2 spectrum with a minimum valley between ions (1 % maximum intensity) to ensure that overlapping signals are detected as separate entities. Next, all extracted ions are filtered to be within a relatively wide tolerance (+ / - 10 mDa for acquisitions in a high-resolution detector or 0.8 mDa for low-resolution) of each expected reporter ion *m/z* for the given quantification method. The reporter ions are then analyzed in a similar way to that described by Pachl[174] to reveal the most coherent set of reporter ions. Because the *m/z* range of reporter ions (5.010 Th) is small, the ions will be equally affected by any calibration offset, but the relative mass difference between them will remain the same. The mass difference between each identified ion and the theoretical *m/z* of the reporter ions – the seed deltas - are stored. Each seed delta is applied to the ions, in turn, to match within a narrow tolerance of 3.16 mDa. The optimum cohort of reporter ions is found that satisfy the following criteria: a) highest count of reporter ions b) highest total reporter ion area and c) lowest spread away from modified *m/z*.

Potential coalescence as described in ref.[170] between proximal TMT11 reporter ion pairs (e.g. TMT127 N & TMT127 C) during acquisition in the Orbitrap (often resulting from high (i.e. > 1e^6) MS^n ion target settings) is estimated by calculating the proportion of intensity overlap between the furthest left (or right) point in the reporter ion peak and the integrated intensity of its counterpart reporter ion. The level of overlap can later be used to filter out data where coalescence has occurred. The values are stored in a dedicated table in the .hdf5 file and reporter ion intensities are multiplied

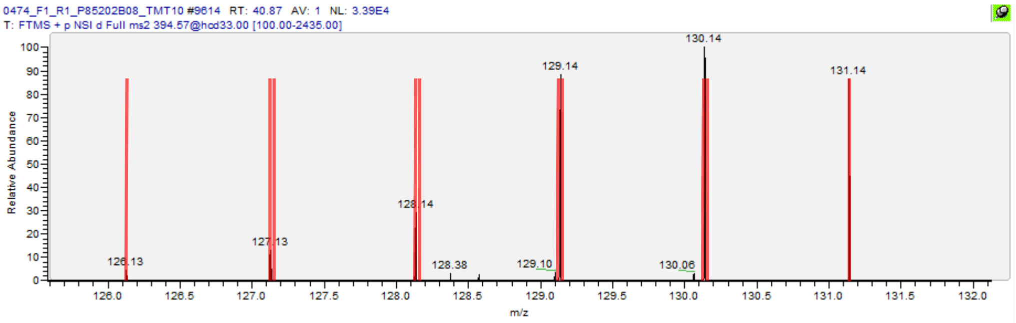


Figure (5): Depiction of expected (red bars) and actual reporter ions (black peaks) present in the low *m/z* range of the MS/MS spectrum for a sample labeled with TMT10-plex reagents. Since the reporter ion masses are known, it is possible to extract the most coherent set of reporter ions from the spectrum even if the acquired masses are offset due to instrument miss-calibration.

with ion accumulation times (the unit is milliseconds) to yield a measure proportional to the number of ions and is referred to here as 'ion area'.

The reporter ion values are then corrected for incorporation of naturally occurring heavy isotope impurities by removing contaminant signals according to values provided by the label manufacturer or from separately performed runs of the individual reporter ions. All quantification data are stored in a dedicated table within the .hdf5 file along with the (MS/MS) spectrum identifier that they were acquired with. This is later used to link these quantification data back to the appropriate data points such as peptides from a database search or the precursor ion.

2.1.5 MS1 signal processing

Once all spectra have been processed as described above, isobarQuant attempts to reassign the precursor *m/z* using chromatographic peak data. XICs are constructed for all ions in the precursor ion isotope cluster. Chromatographic peaks are detected in the XICs and these are linked together to form isotope clusters. The cluster intensity data is compared to the theoretical isotope intensities of an averagine model with similar *m/z*. The cluster with the lowest least squares fit is selected to represent the precursor. The new *m/z* is calculated from the intensity weighted average *m/z* of the top of the monoisotopic peak. The newly-calculated accurate precursor mass is stored in the .hdf5 file with other MS/MS precursor information such as the derived and measured S2I values; summed reporter ions; Full width, half maximum (FWHM) values and area and intensity of the precursor. The bins created during the XIC extraction step are also recorded in the .hdf5 file. Upon completion, indexes are built on key columns in the raw tables. The first step of the pre-Mascot workflow is complete and the .hdf5 file is closed.

2.1.6 Fragment ion deisotoping and deconvolution and .mgf file creation

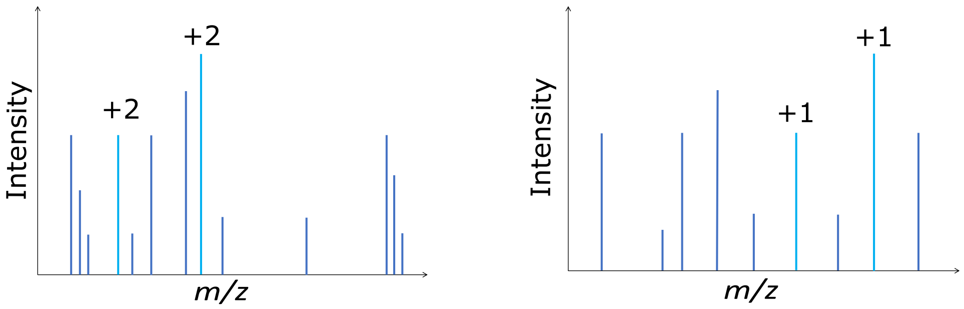
The second and final part of the pre-Mascot workflow is the creation of an input file for Mascot searches. This file type is given the name Mascot Generic Format because it contains data essential for the Mascot search engine (http://www.matrixscience.

com/help/data_file_help.html#GEN), but is meanwhile a general term used for the text-based file format suitable for many search engines and also for manual inspection. MS/MS spectral information is read from .hdf5 file created above and the (fragment) 'ions' table and 'msmsheaders' (selected precursor *m/z*'s) table are queried to extract the appropriate fragment ion and precursor ion information. In experiments quantified using isobaric tags in modes where the MS2 spectrum is used for identification as well as quantification, all reporter ions (determined in 2.1.4) are removed from the fragment ion peak list, since high-intensity reporter ions can adversely affect the Mascot scoring (at least in earlier versions of the search tool) and should be removed (this is also discussed in section 3.1). In addition to this, when data have been acquired in high-resolution, high-accuracy mode, the MS2 spectrum is subject to a deisotoping and deconvolution step so that all ions with multiple charges are re-calculated as singly charged ions. This was shown to improve Mascot scores[165] and spreads the peaks out to a more even distribution[167]. Spectra in this format may be useful for other applications or algorithms that match theoretical spectra to observed peaks. The final step combines the deconvoluted ions into a single ion if the calculated *m/z*'s are within the instrument accuracy of the each another, and where this is the case, the new ion is given an intensity equal to the sum of the intensities of the combined ions and an *m/z* equal to the intensity-weighted mean of the combined ions. The filtered, shifted ions are then written to the .mgf file and the spectrum identifier is stored in the TITLE field for each spectrum. For historical reasons this differs from other common .mgf formats where the MS/MS spectrum identifier is given on a separate line. Recording the spectrum identifier here is essential for downstream workflows which link Mascot results with the raw spectrum they are derived from. The spectrum identifier is also used to map reporter ion values back to the MS/MS spectrum they were acquired in.

For low resolution data, a simple filter is applied that selects the four most intense ions in a given segment of the MS/MS spectrum. The given segment size varies with the charge state of the precursor: spectra from +1 and +2 precursors are segmented every 100 Th and spectra from more highly charged precursors are segmented every 50 Th.

2.1.7 Mascot search

The .mgf files may now be manually submitted to Mascot via the Mascot Daemon (http://www.matrixscience.com/daemon.html). isobarQuant is not currently able to perform this task automatically since it would require a great deal of configuration for each user's individual environment and set up. The Mascot Daemon offers a straightforward and simple way to interface with the Mascot search engine, with simple drag and drop functionality and selection of stored parameter sets for running searches according to user preferences. A smaller helper Python script 'getDatFiles.py' is included in the isobarQuant package and can be called via the Mascot Daemon 'Exter-



(a) Unprocessed MS/MS spectrum prior to deconvolution. The +2 charged fragment ions inhabit the lower half of the *m/z* space and the isotopic clusters of two fragments are visible.

(b) After deconvolution all +2 ions carry a charge of +1 and have moved to a different position within the MS/MS spectrum and the isotopic clusters have been reduced to a single peak

Figure (6): Illustration of result of deconvolution of ions in an MS/MS spectrum

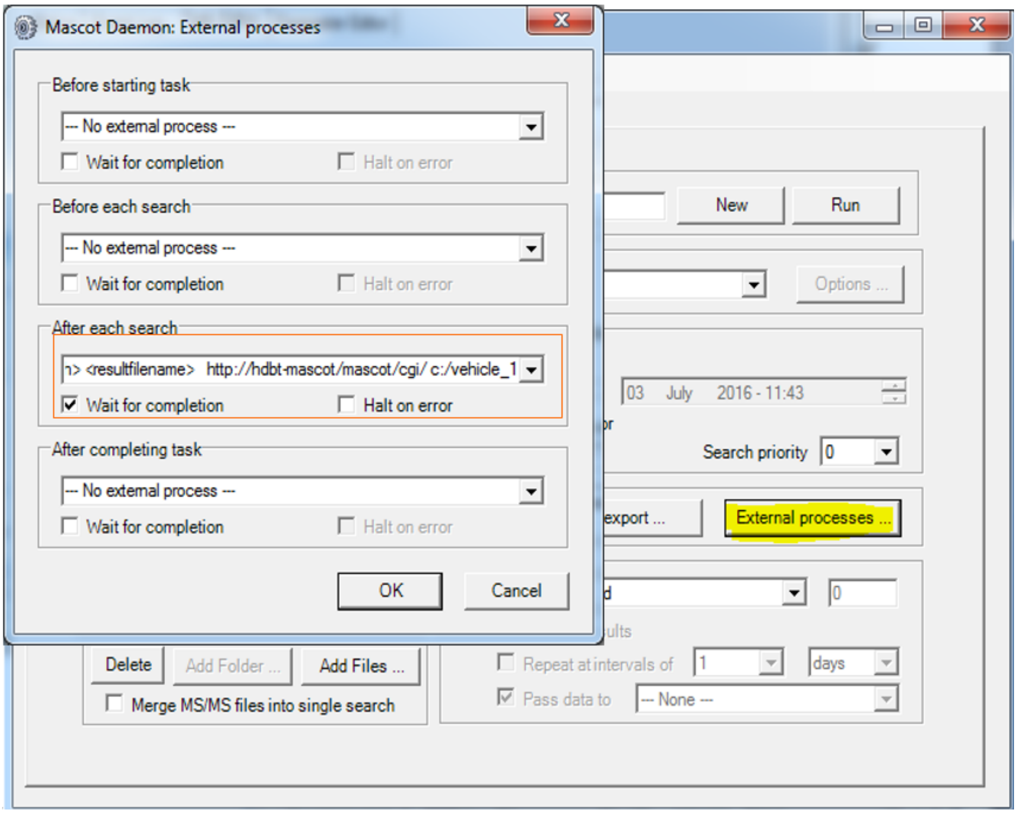
nal processes widget' upon search completion. The user supplied criteria will enable the Mascot daemon to automatically copy the results file to the folder where the pre-Mascot workflow ran. Merging of .mgf files should not be performed at this stage. If an offline 2D-LC pre-fractionation has taken place for which the results of several .raw file acquisitions need to be merged, this merging is performed during the post-Mascot workflow.

2.2 Post-Mascot workflow

As inferred above, the starting point of the post-Mascot workflow is the results of the Mascot search in the form of the automatically copied / downloaded .dat file and the corresponding .hdf5 file created during the pre-Mascot workflow. The raw data can now be stored directly alongside the interpreted peptide and protein information. Like the pre-Mascot workflow the post-Mascot workflow is started via the command line.

2.2.1 Mascot .dat file parsing

The first step of the post-Mascot workflow is to extract pertinent data from the Mascot results file and store this in the .hdf5 file created during the pre-Mascot workflow above. Firstly, the search settings such as fixed and variable modifications, mass tolerances and search database used are recorded and added to the .hdf5 file. The peptide data are QC'd (to remove any peptides containing 'X' amino acids or failing other criteria) and filtered so that only the top candidate peptide (and all other peptides of the same score) and the associated score, modification state, delta to calculated peptide in Dalton and ppm and other similar Mascot-obtained data are stored per MS2 spectrum. The protein associations made by Mascot to these filtered peptides are stored in an additional table within the .hdf5 file where only the single, best scoring, entry per peptide sequence and protein is recorded. This creates a set of data



Mascot Daemon: External processes

Before starting task

--- No external process ---

□ Wait for completion □ Halt on error

Before each search

--- No external process ---

□ Wait for completion □ Halt on error

After each search

ı> <resultfilename> http://hdbt-mascot/mascot/cgi/ c:/vehicle_1

☑ Wait for completion □ Halt on error

After completing task

--- No external process ---

□ Wait for completion □ Halt on error

03 July 2016 - 11:43

Search priority 0

□ Repeat at intervals of 1 days

☑ Pass data to --- None ---

□ Merge MS/MS files into single search

Figure (7): View of Matrix Science Mascot daemon interface exemplifying how to set up the getDatFiles.py script to automatically copy .dat files to a given local directly upon termination of Mascot searches.

with minimal redundancy on which to perform protein inference at a later stage. The presence of high quality peptides can give confidence to protein assignments and may be useful during the protein inference step. A mark of high confidence for a peptide assigned by Mascot, which awards scores based on the inverse log probability of a match arising by chance, is when the score difference between the given match and the next best score is at least ten points. This is equivalent to saying the given peptide is ten times more likely to be correct than the next suggestion. Peptides of lower quality generally have lower scores with much smaller intervals between them. Shorter peptides are also generally less reliable than longer ones for protein inference. isobarQuant has a default requirement that any protein contains at least one of these high-quality peptides with a minimal length of seven amino acids. Such high-confidence peptides are referred to as hook peptides. This value is calculated for all peptides in the .hdf5 file. The .dat file results are stored in the same .hdf5 as the 'raw' data but in a distinct group. It is possible to perform other analyses on these results and append them to this group in order to highlight the dependency on that particular search and its associated parameters. The .hdf5 file is closed and indexes are created on specific columns within the tables.

2.2.2 Protein Inference and peptide FDR calculation

The second step of the post-Mascot workflow is carried out on a single or group of .hdf5 files depending on the mode isobarQuant is running in (merged or unmerged).

One part of this step is the determination of peptide FDR . This is calculated by traversing a list of all identified peptide sequences from all .hdf5 files in descending score order to give a cumulative ratio of decoy to target counts at each score. Peptides mapping to both decoy and target proteins are treated as target peptides. These q-values are stored. Next, isobarQuant (re-)infers protein groups from a good quality subset of the peptides (and their associated proteins) imported from the Mascot .dat files into the individual .hdf5 files during the parsing step. This is achieved by removing those peptides which do not pass the given peptide FDR cut off (usually set to 1 %), removing protein groups and associated peptides that contain only low-quality or repeat peptide identifications and then by applying a principle similar to Occam's razor to create protein groupings based on the remaining shared peptides. Any groups containing completely duplicate (overlapping) peptides are merged together. The peptide groupings are then scanned in descending order of count of hook peptides, total score (sum of all constituent peptides), and then count of peptides passing the FDR threshold. A peptide is marked as 'novel' the first time it is encountered and as non-novel in any groups thereafter. At the end of this scan all groups containing only non-novel peptides are removed. A 'novel' peptide is roughly equivalent to a Mascot bold-red identification. The protein inference step can be performed on one or more .hdf5 files according the requirements of the user via the 'mergeresults' run-time parameter. The results of this part of the processing are recorded in a second .hdf5 file separate from the .hdf5 file created in the pre-Mascot workflow. This ensures that all results (from protein inference and later quantification) are stored in separate files and can summarize the outcome of merging of two or more basic .hdf5 files. The peptide and quantification data is only kept if it is associated with a protein group determined during this stage. The naming convention used for this result .hdf5 file is presented in the Results section (section 3).

2.2.3 Gene level grouping

Grouping peptides solely according to their protein accessions can sometimes result in separate groups which are highly similar but refer to different proteoforms of the same gene. The high number of shared peptides between two groups can, in many cases, lead to few peptides being considered unique and consequently few being used in the quantification of that protein. To circumvent this, and allow researchers to perform comparisons at gene-level, isobarQuant has the possibility to group proteins according to the gene which encoded them. Gene information parsed out of the search results and stored in the .hdf5 proteintable is used to link together identifications of the same gene under a single generated numeric identifier. The generated identifiers and concatenated protein accessions, along with the name and protein descriptions are recorded in the protein table of the results .hdf5 file.

2.2.4 Protein FDR calculation

isobarQuant implements the 'picked protein' approach[130] when calculating the protein-level FDR. isobarQuant assumes that the Mascot searches were performed on a concatenated FASTA database file; containing both decoy and target proteins merged proteins together. This method is preferred because performing separate decoy and target searches can lead to an overly-conservative interpretation of search results[175]. Protein groups are firstly divided into two sets – targets or decoys - according to a suffix in the given accession. If the user has created the decoy proteins using the script provided by Matrix Science (http://www.matrixscience.com/downloads/decoy.pl.gz), all target accessions are prepended with the identifier '###REV###' (or ###RND### for a shuffled [random] decoy database). This decoy 'recognition' sequence used is recorded in the configuration. When a protein set mapping to both the target and decoy of the same accession are identified, the set with the greatest maximum peptide score is picked and the other is discarded. This ensures that a potentially large number of random 'one-hit-wonder' decoy matches do not artificially inflate the count of decoys which would otherwise lead to an overestimation of the protein FDR. Once this picking procedure is complete, the method to calculate the protein FDR is carried out in a way analogous to peptide FDR determination but based on the maximum peptide score per protein group.

2.3 Protein quantification

There are a number of ways to estimate the abundance of proteins present in samples obtained under one or more different conditions, which were discussed in the introduction of this thesis (see 1.4). At this point (and with the first release of the isobarQuant package[161]), the focus is on quantification using isobaric tagging and the 'Top 3'[50] method. An alternative to this would have been to calculate the iBAQ of the proteins as described in ref.[51]. Part III of this report goes into greater detail about the development of isobarQuant for MS1-based chemical labeling strategies.

2.3.1 Peptide quantification: S2I correction

The third step in the post-Mascot workflow is applicable to MS2-based quantification methods and aims to reduce the effect of ratio compression when performing downstream protein quantification. The reporter ion intensities are corrected by a simple algorithm using the signal-to-interference measure, S2I, which has previously been shown to strongly reduce the effect of co-fragmentation and produce more accurate peptide and protein fold changes[171]. Briefly, it assumes that the level of interference derived from the MS1-scan(s) [S2I value] should apply equally to the reporter ion signals coming from co-eluting peptides. This assumption, in combination with the fact that most peptides in the sample are present at similar ratios in all conditions, makes it possible to estimate the amount of signal attributable to the co-eluting peptides and subtract this from each reporter ion. This is done by normalizing the median value of the proportion of total signal for each reporter ion over all peptides to one. The corrected reporter ion intensities are recorded in a different column in the same table as the original reporter ion values to facilitate data checking and QC. The user may decide to leave out this correction step.

2.3.2 Transfer of quantification data

All peptides and associated reporter ions from each .hdf5 file that link to protein sets determined in 2.2.2 are now recorded in the results .hdf5 file. The reporter ion data is kept in a table which also contains information that will later be used to filter them depending on user criteria. This criteria include flags such as uniqueness, P2T, S2I, delta of score from candidate to next Mascot suggestion and q-value. This is the point at which the results of different modes of quantification (i.e. isobaric / MS2 and metabolic / MS1) converge with the protein groups. Because quantification information associated to one protein via peptides potentially located across different .hdf5 files, this part of the process makes intensive use of the file-indexes created in the earlier steps. The quantification of proteins from peptides in many merged .hdf5 files would probably not be possible within dynamic memory on a standard desktop machine.

2.3.3 Performing protein quantification

The fourth processing step in the post-Mascot workflow performs protein quantification. Here the quantification values for all peptides linked to each protein group are extracted from the results .hdf5 file; the filters (as described below) are applied to these reporter ion values and if there are more than a given number of spectra with associated reporter ions, a bootstrapped-sum ratio calculation is carried out as described[173]. The number of bootstrap iterations is set to 5000 and the result has three components: the median fold change over all iterations and the positions of the 0.025 (lower) and 0.975 (upper) quantiles. If the required minimum number of spectra with reporter ions (default is set to 4) is not attained, a simple sum ratio is calculated and the upper and lower quantiles are set to -1. In cases where no quantification data is available in the ‘reference’ channel, no fold change calculation is possible and a value of -1 is recorded (these -1 values are converted later to NA in the text outputs). In cases where no reporter signal is present for that isotope but a signal for the ‘reference’ channel is present then the fold change is set at zero with -1 for the upper and lower quantile values. At this point the area (intensity multiplied by trap fill time) of all reporter ions for each channel is summed and recorded. The number of spectra (PSMs) and unique peptides with associated reporter ions used in the fold change calculation is also noted in the results .hdf5 file.

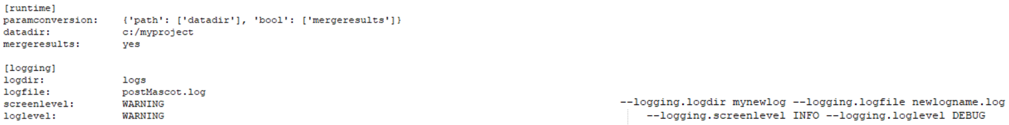
2.3.3.1 Top-3 quantification Taking an average of MS1-based signal intensities has been reported as a good proxy for measuring the absolute protein concentration in a sample[50], or it can at least be used to compare the abundance of a given protein across independent experiments. It is implemented in isobarQuant as part of the post-Mascot workflow. For each protein group, all unique, rank one peptides passing the FDR cut-off are selected, and where applicable additionally filtered to retain those peptides used for MS2 quantification. The XICs (recorded during the pre-Mascot workflow) are extracted for these peptides if they were acquired up to 30s prior to the peak of the XIC. The intensity of the precursor of the best scoring, peptide charge-state modstring (PCM) closest to the peak of the XIC is kept and intensities for different PCMs of the same peptide are summed. Finally, the mean of the $\log_{10}$ intensities of the three most intense ions is recorded and linked to the given protein group within the results .hdf5 file.

2.3.4 Output generation

There are two types of output produced by isobarQuant. The first, as already mentioned in the previous steps, is the .hdf5 file. One .hdf5 file is created and named after each processed .raw file, there is also one .hdf5 file created for each set of results: each time the post-Mascot workflow is run. If isobarQuant is running in merged mode then just one unified results file will be created, otherwise one for each processed .raw file. These binary files are readable through different APIs (such as PyTables or Pandas for python or rh5 in R), but may also be viewed directly using various pure visualization tools such as HFD5View (https://www.hdfgroup.org/downloads/hdfview/) and VITables (http://vitables.org/). The second type of output, the text output, consists of three files (excluding the .mgf files used for Mascot searching): the proteins output, the peptides output and a summary. Again, if the post-Mascot workflow was run in non-merged mode, individual outputs will be created for each .raw file processed. As the name suggests the protein output reports information on the protein level giving fold changes, scores and limited meta information such as protein name and gene name. This output can be the basis for downstream analyses using protein fold changes and associated protein information. The peptides output contains useful information about the peptides associated with the protein groups. It may be useful for peptide-specific analyses or to see which peptides led a protein identification or quantification and which were excluded. The third text-based output is a summary giving basic statistics for the individual samples which are processed during the isobarQuant run.

2.3.5 Multi-threading

isobarQuant is started with a single command line parameter for the pre- and post-Mascot workflows. In both cases a single argument is given pointing to the location of all files to process. In its first implementation the software would traverse this list of files and process one file after another. Since this made inefficient use of available computer resources, a multi-threaded option was added. Here, the user specifies in the configuration how many of the available processors to utilize during the run. The multiprocessing module of Python is then used to set up a mini-queuing system into which all the files are fed and redistributed to different cores. Once the queue



```
[runtime]
paramconversion:    {'path': ['datadir'], 'bool': ['mergeresults']}
datadir:            c:/myproject
mergeresults:       yes

[logging]
logdir:             logs
logfile:            postMascot.log
screenlevel:        WARNING
loglevel:           WARNING
```

```
--logging.logdir mynewlog --logging.logfile newlogname.log
  --logging.screenlevel INFO --logging.loglevel DEBUG
```

Figure (8): Example of configuration file. (Left) 'runtime' and 'logging' in squared brackets are the section headers with text before the colon representing the parameter and after it, the value of that parameter. Changing this value and saving the file will affect all runs performed after the change. The second output (right) shows the same configuration as stored in the configuration file but supplied via the command line (typical for use in a one-off run). The section and parameter are merged using a dot and preceded with two dashes ('- -'). The value for the given parameter is given after the space.

of files to process is worked off, isobarQuant returns to single central processing unit (CPU) mode to continue the processing. The number of CPUs in use can be modulated to 'all' or only some of the number available; this would be useful if isobarQuant is running on a desktop PC that is also being used for other, day-to-day tasks. This leads to a substantial speed up in processing of file on multiprocessor machines.

2.3.6 Configuration

isobarQuant is designed to be simple to use and run 'out of the box' for as many different applications as possible. When necessary, configuration of the system to run with parameters other than those set as 'default' settings is done by editing small text-based configuration files.These follow the standard Python configuration convention in that they are organized in sections which are separated by section headers (indicated by squared brackets). If the parameter change is due to be permanent the configuration file should be updated and save accordingly. If however a one-off change is sufficient the syntax shown in fig. 8 may be used to temporarily overwrite the stored parameter.

2.4 Comparison to MaxQuant using *E. coli* dilution series spiked into human background

At the time of development and publishing of isobarQuant, no software tool was freely available to perform isobaric quantification including and certainly none were available to mitigate the effects of ratio compression, hence no bench marking against the performance of other software was performed at the time. As part of this report, however, a comparison against MaxQuant was carried out. The latest (December 2019) version of this software (MaxQuant 1.6.10.43) was downloaded from https://www.maxquant.org/download_asset/maxquant/latest and used to process the same raw file as isobarQuant.

2.4.1 Sample preparation

The comparison was made by running the same file through isobarQuant and MaxQuant and looking at the difference in peptide quantification values. The raw file used for processing contained data from a dilution series of known concentrations of *E. coli* proteins either spiked into a background of human proteins (labeled and mixed at a ratio of 1:1 for all labels) or with no human background.

E. coli proteins were cultured and purified as follows: 50 µL of DH5α competent *E. coli* cells (18265-017, Invitrogen, stored at -80C) and plasmid (Reference: EP02538, D3729 (2), JMJD3 (1142-1682) H1390A; pcDNA3 N-Flag, prepared by dilution of 1µL in 999µL sterile water to yield (0.8µg/µl)) were thawed on ice. *E. coli* cells were resuspended, and 6.25µl plasmid was added. The sample was mixed gently and left on ice for 30 minutes. Next, the sample was placed for 40 seconds in a 42°C water bath and then put on ice for two minutes. 800µl pre-warmed (37°C) super optimal broth with catabolite repression (SOC) medium was added and the sample incubated for 1 hour at 37°C with 850 rpm rotation. Finally the sample was transformed overnight on lysogeny broth (LB) plates to confer ampicillin resistance. To grow the transformed *E. coli* cells, a 5L erlenmeyer flask with chicanes was filled with 300 ml LB medium containing 100µg/ml ampicillin and placed in an incubator at 37°C for 20 minutes and then incubated overnight at 37°C under rotation (180 rpm). To ensure bacteria were harvested during the exponential growth phase, they were incubated for a further three hours following 20 fold dilution with LB-medium when they reached an OD of 600nm. Cells were harvested by spinning 15 minutes at 3500 rpm, 4°C. Supernatant was aspirated and cell pellets were suspended in phosphate-buffered saline (PBS) before spinning again for 15 min at 3500 rpm, after which they were resuspended in 10ml PBS following supernatant removal. Cells were lyzed using 450µl lysis buffer containing 4% sodium dodecyl sulfate (SDS) and were placed in a thermomixer (Thermo Fisher Scientific) for 3 minutes at 95°C. The sample's SDS concentration was reduced to 2% by addition of 900µl 50mM Tris and was then treated with 62µl benzonase solution (Sigma E1014-25KU) with subsequent incubation at 37°C and 800 rpm for 30 minutes. A second round of incubation under the same conditions as the first, using with half the amount of benzonase solution was carried out for 45 minutes. Lysates were cleared by centrifugation (20,000xg) for 20 minutes at room temperature and supernatent transfer to fresh tubes. Protein amount, determined by bicinchoninic acid (BCA) assay was 0.1µg/µL.

Four Human cell lines (HEK293, K-562, HepG2 and placenta) were cultured inhouse on medium over three days in standard conditions and harvested by spinning at 3500 rpm, 4°C and then lysed in buffer containing 4% SDS, with subsequent clearing via centrifugation (20,000xg) for 20 minutes at room temperature. The cell line lysates were mixed in a 1:1 ratio and final protein amount was determined by BCA to be 0.56µg/µL.

2.4.1.1 Sample preparation for mass spectrometry and TMT-labeling *E. coli* samples were divided into nine equal aliquots, human samples into six, with each one being labeled with a different TMT reagent (see 2). Following reduction by dithiothreitol (DTT) and alkylation with iodacetamidem, samples are prepared for MS analysis via a gel-free, SP2 approach, using hydrophilic beads to bind proteins, remove contaminants and subsequently perform protease digests; the resulting peptides were eluted

	126	127N	127C	128N	128C	129N	129C	130N	130C	131
E. coli	0	1	1	0.5	0.5	0.2	0.2	0.1	0.1	1
Human	1	1	0	1	0	1	0	1	0	1

Table (2): Description of dilution series for TMT reporter ions used in comparison between MaxQuant and isobarQuant

from the beads and TMT-labeled. This approach is based on trapping of proteins on the surface of hydrophilic carboxy-functionalized magnetic beads in excess of organic solvent (via HILIC) as described by Hughes *et al.*[176], but peptides were trapped in 96-well filter plates rather than via magnetic separation. Proteins were digested with trypsin and Lys-C overnight and were re-suspended in 10µL water. 10µL TMT reagent was added and incubated for 1 hour at room temperature with shaking at 500 rpm. Individual reactions were quenched using 2.5% NH_2OH in 0.1 M HEPES and then pooled into a single tube, washed with buffer and subsequently lyophelized.

The final *E. coli* ratios were obtained by diluting above labeled samples with buffer to a final volume as given in table 2.

Following the digestion and labeling steps described above, the sample was acquired using a gradient of 115 minutes. Samples were dried *in vacuo* and re-suspended in 0.05 %trifluoroacetic acid (TFA) in water. Of the sample, 50% was injected into an Ultimate3000 nanoRLSC (Dionex, Sunnyvale, CA) coupled to a QExactive HF-X (Thermo Fisher Scientific). Peptides were trapped on a 5mm x 300 µm C18 column (Pepmap100, 5 µm, 300 Å, Thermo Fisher Scientific) in water with 0.05 % TFA at 60 °C. Separation was performed on custom 50 cm × 100 µM (ID) reversed-phase columns (Reprosil) at 55°C. Gradient elution was performed from 2% acetonitrile to 40% acetonitrile in 0.1% formic acid and 3.5% dimethyl sulfoxide (DMSO) over 105 minutes. Samples were online injected into a QExactive HF-X mass spectrometer operating with a data-dependent top 15 method. MS spectra were acquired using 60,000 resolution and an ion target of $3x10^6$. Higher energy collisional dissociation (HCD) scans were performed with 31% normalized collision energy (nCE) at 30 000 resolution (at *m/z* 200), and the ion target setting was fixed at $2x10^5$. The instruments were operated with Tune 3.1 and Xcalibur 4.0.27.10.

2.4.2 Data Processing

2.4.2.1 MaxQuant MaxQuant was set up to process the raw files as follows: Protein identification and quantification was performed using MaxQuant v 1.6.2.3 using Andromeda as the search engine. A downloaded version of Uniprot Human and 2018 was applied for matching MS/MS spectra. TMT10 quantification of peptide and protein abundances was used. Cysteine carbamidomethylation was used as a fixed modification; methionine oxidation and acetylation at protein N-termini were used as variable modifications for both identification and quantification. Trypsin/P was selected as enzyme specificity with maximum of three missed cleavages allowed. 1% false discovery rate was used as a filter at both protein and peptide levels. The inbuilt contaminant database was selected and 'reversed protein' was selected to create decoys. The correction factors (for isotope impurity correction) used were those supplied by the manufacturer for the corresponding batch of TMT-reagents. All other parameters were left at their default setting.

To determine peptide fold changes the peptides.txt, evidence.txt and msms.txt file outputs of MaxQuant were interrogated and joined in a Python Jupyter notebook. Peptide fold change ratios were determined by dividing the value in the 'Reporter intensity corrected' column for each TMT label by the intensity-corrected column corresponding to TMT127C (baseline dilution of *E. coli* sample with no human background). Unless otherwise stated, only fold changes of unique *E. coli* peptides with a parent ion fraction (PIF) >0.75 were used for analyses.

2.4.2.2 isobarQuant isobarQuant was used with the settings described in the sections above. The file supplied to Mascot 2.5.1 for MS/MS searching was essentially identical to that used for Andromeda, except that it consisted of a pre-concatenated file comprising Uniprot Human and *E. coli* (downloaded December 2018) together with their reverse protein counterparts as decoys. The Mascot search settings were carbamidomethylation of cysteine and TMT6plex of lysine as fixed modifications; acetylation of protein N-termini, oxidation of methionine and modification of TMT6plex on peptide N-termini as variable modifications. Trypsin/P was selected as enzyme specificity with maximum of three missed cleavages allowed and a precursor tolerance of 20ppm and fragment ion tolerance of 0.02 Da.

The correction factors used in isobarQuant were based on measured values and therefore do not just take the isotope impurities of the individual carbon (or nitrogen) atoms used during manufacture into account, but also incorporate any potential additive effects and cross talk between them. These were determined by measurement of single samples, each labeled with one of the ten individual TMT reporter ions. The proportion of signal leading to cross-talk in channels other than that used for labeling the sample is then easy to calculate. Where this measured proportion was greater than 1%, it was set to be used by isobarQuant. These determined values are given in table 3 and are supplied as defaults with the isobarQuant software.

To determine peptide fold changes, the specquant table in the 'result.' hdf5 output was interrogated directly in a Python Jupyter notebook. Peptide fold change ratios were determined by dividing the value in the 'quant_all_corrected' column for each TMT label by the intensity-corrected column corresponding to TMT127C (baseline dilution of *E. coli* sample with no human background). Only fold changes of unique *E. coli* peptides with an S2I >0.75 and otherwise fulfilling all isobarQuant default peptide quantification criteria (P2T >4, FDR<1%, Mascot score >15 and peptide length > 6 amino acids) were analyzed.

	126	127N	127C	128N	128C	129N	129C	130N	130C	131
126	x		0.04							
127N		x		0.0382						
127C			x		0.0335					
128N			0.0168	x		0.0351				
128C			0.0148		x		0.0286			
129N				0.0177		x		0.0279		
129C					0.0149		x		0.0234	
130N						0.0265	0.0175	x		0.0234
130C							0.0329		x	
131								0.032		x

Table (3): TMT10 correction factors. Shown here are the default values for TMT10 isotope correction supplied with isobarQuant. The rows represent the individual reporter ion channels and the amount of isotope impurity identified in other channels is given as a fraction of the total reporter signal for the values in the row. They are based on actual measurements of individual TMT10 labels. These values will differ slightly with each separate batch of TMT reagent.

2.5 Mitigation of ratio-compression for published dataset

As part of their study into plasma proteins Keshishian *et al.*[177] used a spike-in of 97 synthesized SILAC heavy peptides labeled in a dilution series in order to measure ratio compression and across 30 patient samples. The heavy peptides were labeled with either iTRAQ or TMT-6plex or TMT-10plex in several dilution series. This enabled the authors to assess the level of ratio compression for all three labels by comparing the obtained ratios of the peptide fold changes, acquired against a high background, with the expected values and entirely without interfering with the experiment itself. The dataset comprising of 30 raw files labeled with TMT10 was downloaded from the authors' repository ftp://massive.ucsd.edu/MSV000079033/raw/iTRAQ_TMT_Comparison/Plasma_TMT10/ and processed with isobarQuant as described above in 'mergeresults' mode. Mascot searches were performed with the standard settings as above: carbamidomethylation of cysteine as a fixed modification and acetylation of protein N-termini, oxidation of methionine and TMT6plex on peptide N-termini as variable modifications, but also included a fixed modification for a heavy SILAC modification on arginine residues and the fixed TMT6plex modification on lysine was increased to incorporate the mass of heavy SILAC. isobarQuant was slightly modified at this point to ensure that the reporter ion signals from all acquired MS/MS spectra were used in the calculation of the ratio compression correction factors, not just those with a Mascot peptide hit. This was necessary because the vast majority of peptides in the dataset are from the plasma protein sample (which contribute to the non-specific background TMT reporter ion signals) using only the values from the heavy peptides would be counter-intuitive. Following processing with isobarQuant, the peptides output of workflow was loaded into a Jupyter notebook and used to calculate peptide fold changes against label 128C as the baseline (as in the original publication). The post-Mascot pipeline of isobarQuant was run a second time using the 30 downloaded files processed in the first round, but this time the parameter 'run_corrects2iquant' was set to 'no' in the postMascot.cfg file to turn off the S2I correction. This allows the direct examination of the effect of isobarQuant's ratio compression correction.

3 Results

The main goal of isobarQuant project was to create a simple, open source, standalone command line tool that could encompass the experiences and algorithmic developments of many experimenters over previous years to perform isobaric quantification on one or more (merged) datasets. It should be easy to install and to use. The tool should adhere to the 'rule of thumb' in computational mass spectrometry that the processing of the data should not take longer than original acquisition. The application should be equally at home on a laptop / user's desktop or on a larger server and the outputs it creates should be easily queryable and amiable to downstream applications.

isobarQuant fulfills these aims and lies at the heart of the thermal proteome profiling (TPP) analysis described in Franken and Mathieson *et al.*[161], it is available on GitHub via the URL https://github.com/protcode/isob.git or as a zip file https://github.com/protcode/isob/archive/master.zip. Its installation and running is very simple and, following download, can be described in much less than half a page of text and in fewer than six steps if Python is already installed (see https://github.com/protcode/isob/blob/master/QuickStartGuide.pdf).

A more detailed discussion of how isobarQuant was further developed and its results can be used as the basis for the next two parts of this work (parts III & IV).

For each isobarQuant run the number of output files will depend on the number of input (.raw) files and whether or not the user requires file merging. For each .raw file an .hdf5 file and .mgf file will be created. Each of the .mgf files submitted to Mascot will yield one .dat file, which is then internalized into its corresponding .hdf5 file. The results of the protein inference and quantification will yield at least four further files (one results .hdf5 file and three .txt outputs). Whilst this may seem to negate the original aim of having all data in one single file, it is worth stating that the additional results .hdf5 file is just an aggregation of data contained in each of the original .hdf5 files and the .txt outputs are available for the convenience of the user to be able browse the data or for downstream process that do not have any .hdf5 API.

The name of the output files are named according to the following convention:

<directory>_<analysis_type>_<rundate> _<runtime>_[output _type]

- directory is replaced by the first 25 characters of the data directory name; analysis_type is 'merged_results' when a merged analysis has been performed and is the .raw file name plus '_results' when each .raw file is analyzed separately
- run date is the date when the processing was carried out in the format YYYYMMDD
- run time is the time when the process was run in format HHMM
- output_type is either
  1. _proteins.txt: file contains protein information and corresponding protein fold changes. This file can be taken to the next stage.

2. _peptides.txt: file contains information about the individual peptides and their reporter ion values.
3. _summary.txt: file contains some statistical information about the runs performed
4. .hdf5: the underlying .hdf5 file contains the results of the run started above

The filename convention was chosen to include the parent folder name within it since the folder name is often related to the experiment being analyzed. The inclusion of the date/time stamp in the file name prevents the accidental overwriting of data, especially if multiple operating parameters are being investigated.

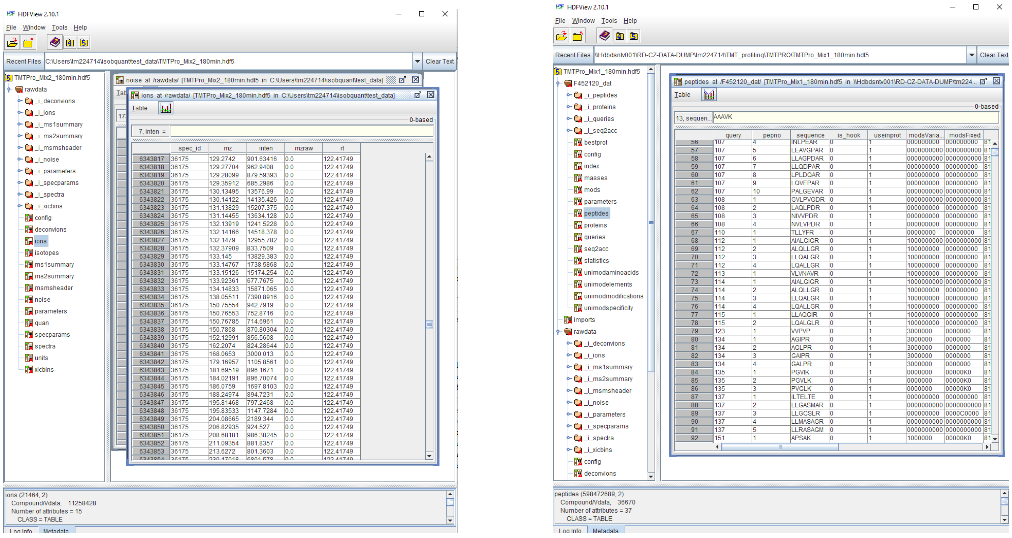
For ease of interpretation and for use in downstream analyses, the protein output of isobarQuant includes gene information so that proteins can be easily grouped and the protein FDR is also included so that proteins can be further filtered if required. The peptide output also includes a lot of detailed information concerning the acquisition and identification of the peptide and whether or not it was used for quantification or in the determination of the protein group.

The structure of all .hdf5 file outputs enables quick access to data in an encapsulated way and the indexing of the .hdf5 file in tandem with a Python API allow fast access and in-kernel selection of data to reduce the memory footprint. The .hdf5 file may be accessed directly by any programming languages that have a suitable API. R, for example has at least two .hdf5 readers (rhdf5 in the BioConductor Suite or h5 in the CRAN repository) and may be of particular interest for bioinformatic applications. Identification and quantification information are stored parallel to the raw spectral data and the parameters used for processing in a single, self-contained, file. A separate results file is created each time the post-Mascot workflow runs. If result data (Mascot .dat information) is already contained in the .hdf5 file at the time of running the post-Mascot workflow, the user is prompted whether the program should overwrite the existing results with new data.

The whole isobarQuant package is published as an open-source project completely free of charge. isobarQuant is started via the command line, allowing it to be easily incorporated into other workflows if required. It locates files to process according to user-supplied parameters. On a 4-processor desktop machine it can run overnight to completely process more than ten TMT-quantified files.

3.1 isobarQuant for concomitant interrogation of multiple .hdf5 files

The .hdf5 file format allows fast access to the underlying data from one or more .hdf5 files. Figure 11 shows a Python / Pandas code snippet highlighting the ease with which one can extract data for, as an example, building a consensus spectral library. This process took approximately 100s to extract the required data from ~300 .hdf5 files on a SATA file system. The timing will, of course, depend on the speed of the file system,



C:\Users\tm224714\isobquant\test_data\TMTPro_Mix2_180min.hdf5

TMTPro_Mix2_180min.hdf5
- rawdata
  - _i_deconvions
  - _i_ions
  - _i_ms1summary
  - _i_ms2summary
  - _i_msmsheader
  - _i_noise
  - _i_parameters
  - _i_specparams
  - _i_spectra
  - _i_xicbins
  - config
  - deconvions
  - ions
  - isotopes
  - ms1summary
  - ms2summary
  - msmsheader
  - noise
  - parameters
  - quan
  - specparams
  - spectra
  - units
  - xicbins

noise at /rawdata/ [TMTPro_Mix2_180min.hdf5 in C:\Users\tm224714\isobquant\test_data]

ions at /rawdata/ [TMTPro_Mix2_180min.hdf5 in C:\Users\tm224714\isobquant\test_data]

7, inten =

	spec_id	mz	inten	mzraw	rt
6343817	36175	129.2742	901.63416	0.0	122.41749
6343818	36175	129.27704	962.9408	0.0	122.41749
6343819	36175	129.28099	879.59393	0.0	122.41749
6343820	36175	129.35912	685.2986	0.0	122.41749
6343821	36175	130.13495	13576.99	0.0	122.41749
6343822	36175	130.14122	14135.426	0.0	122.41749
6343823	36175	131.13829	15207.375	0.0	122.41749
6343824	36175	131.14455	13634.128	0.0	122.41749
6343825	36175	132.13919	1241.5228	0.0	122.41749
6343826	36175	132.14166	14518.378	0.0	122.41749
6343827	36175	132.1479	12955.782	0.0	122.41749
6343828	36175	132.37909	833.7509	0.0	122.41749
6343829	36175	133.145	13829.383	0.0	122.41749
6343830	36175	133.14767	1738.5868	0.0	122.41749
6343831	36175	133.15126	15174.254	0.0	122.41749
6343832	36175	133.92361	677.7675	0.0	122.41749
6343833	36175	134.14833	15871.065	0.0	122.41749
6343834	36175	138.05511	7390.8916	0.0	122.41749
6343835	36175	150.75554	942.7919	0.0	122.41749
6343836	36175	150.76553	752.8716	0.0	122.41749
6343837	36175	150.76785	714.6961	0.0	122.41749
6343838	36175	150.7868	870.80304	0.0	122.41749
6343839	36175	152.12991	856.5608	0.0	122.41749
6343840	36175	162.2074	824.28644	0.0	122.41749
6343841	36175	168.0653	3000.013	0.0	122.41749
6343842	36175	179.16957	1105.8561	0.0	122.41749
6343843	36175	181.69519	896.1671	0.0	122.41749
6343844	36175	184.02191	896.70074	0.0	122.41749
6343845	36175	186.0759	1697.8103	0.0	122.41749
6343846	36175	188.24974	894.7231	0.0	122.41749
6343847	36175	195.81468	797.2468	0.0	122.41749
6343848	36175	195.83533	1147.7284	0.0	122.41749
6343849	36175	204.08665	2189.344	0.0	122.41749
6343850	36175	206.82935	924.527	0.0	122.41749
6343851	36175	208.68181	986.38245	0.0	122.41749
6343852	36175	211.09354	881.8357	0.0	122.41749
6343853	36175	213.6272	801.3603	0.0	122.41749
6343854	36175	220.17018	6801.578	0.0	122.41749

ions (21464, 2)
Compound/Vdata, 11258428
Number of attributes = 15
CLASS = TABLE

\\Hdbdsntv001\RD-CZ-DATA-DUMP\tm224714\TMT_profiling\TMTPRO\TMTPro_Mix1_180min.hdf5

TMTPro_Mix1_180min.hdf5
- F452120_dat
  - _i_peptides
  - _i_proteins
  - _i_queries
  - _i_seq2acc
  - bestprot
  - config
  - index
  - masses
  - mods
  - parameters
  - peptides
  - proteins
  - queries
  - seq2acc
  - statistics
  - unimodaminoacids
  - unimodelements
  - unimodmodifications
  - unimodspecificity
- imports
- rawdata
  - _i_deconvions
  - _i_ions
  - _i_ms1summary
  - _i_ms2summary
  - _i_msmsheader
  - _i_noise
  - _i_parameters
  - _i_specparams
  - _i_spectra
  - _i_xicbins
  - config
  - deconvions

peptides at /F452120_dat/ [TMTPro_Mix1_180min.hdf5 in \\Hdbdsntv001\RD-CZ-DATA-DUMP\tm224...

13, sequen... AAAVK

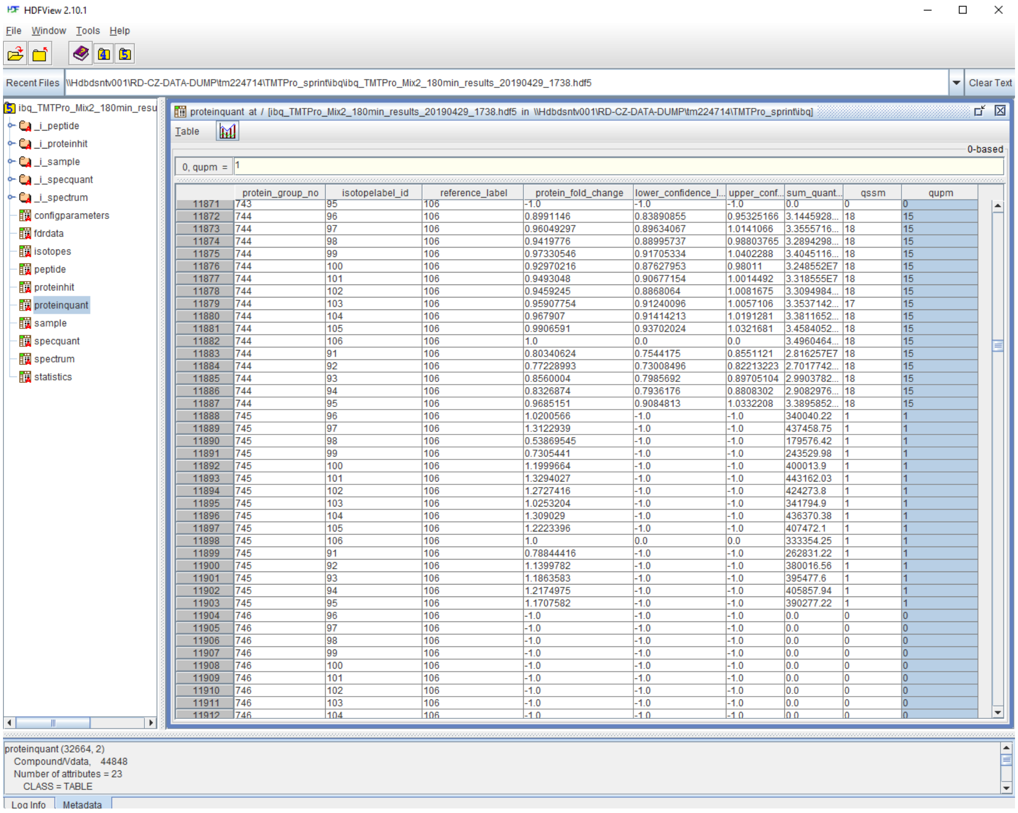
	query	pepno	sequence	is_hook	useinprot	modsVaria...	modsFixed	
56	107	4	INLPEAR	0	1	000000000	000000000	81
57	107	5	LEAVGPAR	0	1	0000000000	0000000000	81
58	107	6	LLAGPDAR	0	1	0000000000	0000000000	81
59	107	7	LLQDPAR	0	1	000000000	000000000	81
60	107	8	LPLDQAR	0	1	000000000	000000000	81
61	107	9	LQVEPAR	0	1	000000000	000000000	81
62	107	10	PALGEVAR	0	1	0000000000	0000000000	81
63	108	1	GVLPVGDR	0	1	0000000000	0000000000	81
64	108	2	LAQLPDR	0	1	000000000	000000000	81
65	108	3	NIVVPDR	0	1	000000000	000000000	81
66	108	4	NVLVPDR	0	1	000000000	000000000	81
67	110	1	TLLYFR	0	1	00000000	00000000	81
68	112	1	AIALGIGR	0	1	1000000000	0000000000	81
69	112	2	ALQLLGR	0	1	100000000	000000000	81
70	112	3	LLQALGR	0	1	100000000	000000000	81
71	112	4	LQALLGR	0	1	100000000	000000000	81
72	113	1	VLVNAVR	0	1	100000000	000000000	81
73	114	1	AIALGIGR	0	1	1000000000	0000000000	81
74	114	2	ALQLLGR	0	1	100000000	000000000	81
75	114	3	LLQALGR	0	1	100000000	000000000	81
76	114	4	LQALLGR	0	1	100000000	000000000	81
77	115	1	LLAQGIR	0	1	100000000	000000000	81
78	115	2	LQALGLR	0	1	100000000	000000000	81
79	123	1	VVPVP	0	1	3000000	0000000	81
80	134	1	AGIPR	0	1	3000000	0000000	81
81	134	2	AGLPR	0	1	3000000	0000000	81
82	134	3	GAIPR	0	1	3000000	0000000	81
83	134	4	GALPR	0	1	3000000	0000000	81
84	135	1	PGVIK	0	1	0000000	00000K0	81
85	135	2	PGVLK	0	1	0000000	00000K0	81
86	135	3	PVGLK	0	1	0000000	00000K0	81
87	137	1	ILTELTE	0	1	000000000	000000000	81
88	137	2	LLGASMAR	0	1	0000000000	0000000000	81
89	137	3	LLGCSLR	0	1	000000000	0000C0000	81
90	137	4	LLMASAGR	0	1	0000000000	0000000000	81
91	137	5	LLRASAGM	0	1	0000000000	0000000000	81
92	151	1	APSAK	0	1	1000000	00000K0	81

peptides (598472689, 2)
Compound/Vdata, 36670
Number of attributes = 37
CLASS = TABLE

(a) Snapshot of .hdf5 file showing the results of the pre-Mascot workflow. On the left is the list of tables within the file, on the right an excerpt of the ions table (where all ions from all MS/MS spectra are recorded).

(b) Snapshot of the .hdf5 file focusing on the results of the post-Mascot workflow. The hierarchical structure is visible on the left side. The data in table 'peptides' can be seen on the right where all Mascot-assigned ranked peptides along with some of their associated values, such as modifications and score are given. At the top one can see the indexes associated with the file which aid in data selection and access.

Figure (9): The .hdf5 file after the pre- and post-Mascot workflows.



\\Hdbdsntv001\RD-CZ-DATA-DUMP\tm224714\TMTPro_sprint\ibq\ibq_TMTPro_Mix2_180min_results_20190429_1738.hdf5

ibq_TMTPro_Mix2_180min_resu
- _i_peptide
- _i_proteinhit
- _i_sample
- _i_specquant
- _i_spectrum
- configparameters
- fdrdata
- isotopes
- peptide
- proteinhit
- proteinquant
- sample
- specquant
- spectrum
- statistics

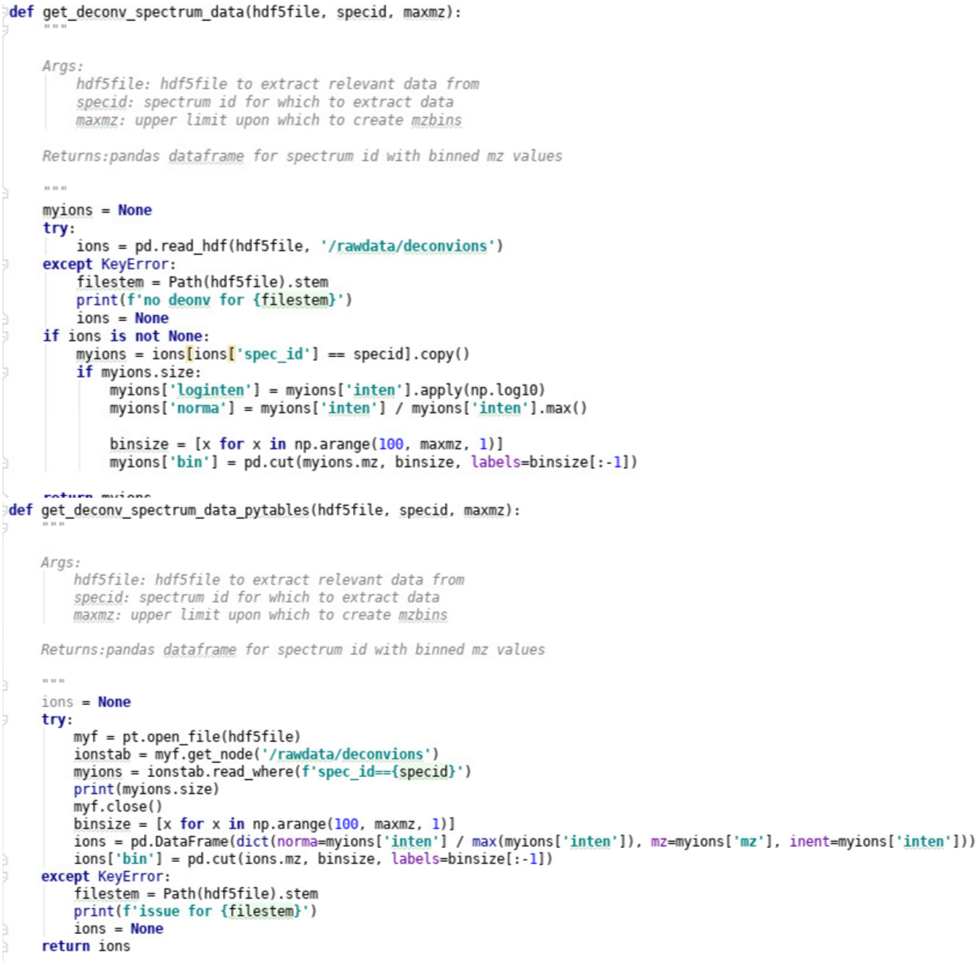
proteinquant at / [ibq_TMTPro_Mix2_180min_results_20190429_1738.hdf5 in \\Hdbdsntv001\RD-CZ-DATA-DUMP\tm224714\TMTPro_sprint\ibq]

0, qupm = 1

	protein_group_no	isotopelabel_id	reference_label	protein_fold_change	lower_confidence_l...	upper_conf...	sum_quant...	qssm	qupm
11871	743	95	106	-1.0	-1.0	-1.0	0.0	0	0
11872	744	96	106	0.8991146	0.83890855	0.95325166	3.1445928...	18	15
11873	744	97	106	0.96049297	0.89634067	1.0141066	3.3555716...	18	15
11874	744	98	106	0.9419776	0.88995737	0.98803765	3.2894298...	18	15
11875	744	99	106	0.97330546	0.91705334	1.0402288	3.4045116...	18	15
11876	744	100	106	0.92970216	0.87627953	0.98011	3.248552E7	18	15
11877	744	101	106	0.9493048	0.90677154	1.0014492	3.318555E7	18	15
11878	744	102	106	0.9459245	0.8868064	1.0081675	3.3094984...	18	15
11879	744	103	106	0.95907754	0.91240096	1.0057106	3.3537142...	17	15
11880	744	104	106	0.967907	0.91414213	1.0191281	3.3811652...	18	15
11881	744	105	106	0.9906591	0.93702024	1.0321681	3.4584052...	18	15
11882	744	106	106	1.0	0.0	0.0	3.4960464...	18	15
11883	744	91	106	0.80340624	0.7544175	0.8551121	2.816257E7	18	15
11884	744	92	106	0.77228993	0.73008496	0.82213223	2.7017742...	18	15
11885	744	93	106	0.8560004	0.7985692	0.89705104	2.9903782...	18	15
11886	744	94	106	0.8326874	0.7936176	0.8808302	2.9082976...	18	15
11887	744	95	106	0.9685151	0.9084813	1.0332208	3.3895852...	18	15
11888	745	96	106	1.0200566	-1.0	-1.0	340040.22	1	1
11889	745	97	106	1.3122939	-1.0	-1.0	437458.75	1	1
11890	745	98	106	0.53869545	-1.0	-1.0	179576.42	1	1
11891	745	99	106	0.7305441	-1.0	-1.0	243529.98	1	1
11892	745	100	106	1.1999664	-1.0	-1.0	400013.9	1	1
11893	745	101	106	1.3294027	-1.0	-1.0	443162.03	1	1
11894	745	102	106	1.2727416	-1.0	-1.0	424273.8	1	1
11895	745	103	106	1.0253204	-1.0	-1.0	341794.9	1	1
11896	745	104	106	1.309029	-1.0	-1.0	436370.38	1	1
11897	745	105	106	1.2223396	-1.0	-1.0	407472.1	1	1
11898	745	106	106	1.0	0.0	0.0	333354.25	1	1
11899	745	91	106	0.78844416	-1.0	-1.0	262831.22	1	1
11900	745	92	106	1.1399782	-1.0	-1.0	380016.56	1	1
11901	745	93	106	1.1863583	-1.0	-1.0	395477.6	1	1
11902	745	94	106	1.2174975	-1.0	-1.0	405857.94	1	1
11903	745	95	106	1.1707582	-1.0	-1.0	390277.22	1	1
11904	746	96	106	-1.0	-1.0	-1.0	0.0	0	0
11905	746	97	106	-1.0	-1.0	-1.0	0.0	0	0
11906	746	98	106	-1.0	-1.0	-1.0	0.0	0	0
11907	746	99	106	-1.0	-1.0	-1.0	0.0	0	0
11908	746	100	106	-1.0	-1.0	-1.0	0.0	0	0
11909	746	101	106	-1.0	-1.0	-1.0	0.0	0	0
11910	746	102	106	-1.0	-1.0	-1.0	0.0	0	0
11911	746	103	106	-1.0	-1.0	-1.0	0.0	0	0
11912	746	104	106	-1.0	-1.0	-1.0	0.0	0	0

proteinquant (32664, 2)
Compound/Vdata, 44848
Number of attributes = 23
CLASS = TABLE

Figure (10): Snapshot of the results .hdf5 file focusing on the proteinquant table at the end of the post-Mascot workflow. The long format shows the protein fold changes, upper and lower confidences and the summed quantification values for 3-4 protein groups. The number of quantified spectra for each channel and quantified unique peptides is also visible on the right-hand side.The reference label and reporter ion isotope label is encoded by a numeric identifier (set in the configuration file) in two further columns.



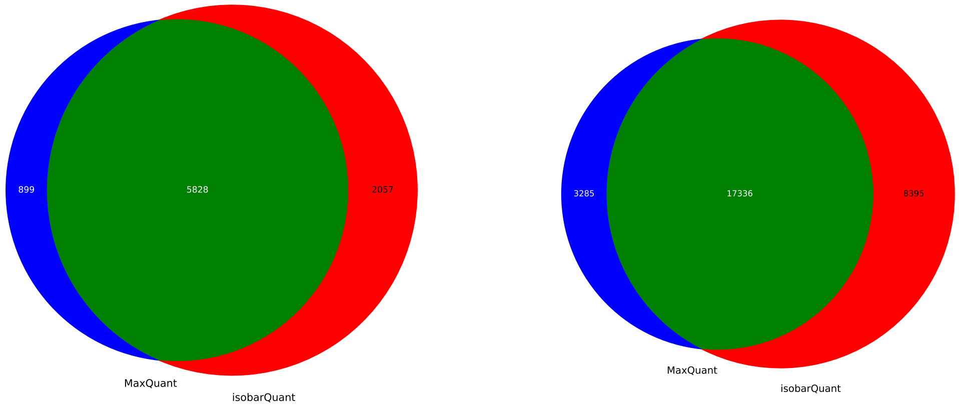
```
def get_deconv_spectrum_data(hdf5file, specid, maxmz):
    """

    Args:
        hdf5file: hdf5file to extract relevant data from
        specid: spectrum id for which to extract data
        maxmz: upper limit upon which to create mzbins

    Returns:pandas dataframe for spectrum id with binned mz values

    """
    myions = None
    try:
        ions = pd.read_hdf(hdf5file, '/rawdata/deconvions')
    except KeyError:
        filestem = Path(hdf5file).stem
        print(f'no deonv for {filestem}')
        ions = None
    if ions is not None:
        myions = ions[ions['spec_id'] == specid].copy()
        if myions.size:
            myions['loginten'] = myions['inten'].apply(np.log10)
            myions['norma'] = myions['inten'] / myions['inten'].max()

            binsize = [x for x in np.arange(100, maxmz, 1)]
            myions['bin'] = pd.cut(myions.mz, binsize, labels=binsize[:-1])

    return myions
def get_deconv_spectrum_data_pytables(hdf5file, specid, maxmz):
    """

    Args:
        hdf5file: hdf5file to extract relevant data from
        specid: spectrum id for which to extract data
        maxmz: upper limit upon which to create mzbins

    Returns:pandas dataframe for spectrum id with binned mz values

    """
    ions = None
    try:
        myf = pt.open_file(hdf5file)
        ionstab = myf.get_node('/rawdata/deconvions')
        myions = ionstab.read_where(f'spec_id=={specid}')
        print(myions.size)
        myf.close()
        binsize = [x for x in np.arange(100, maxmz, 1)]
        ions = pd.DataFrame(dict(norma=myions['inten'] / max(myions['inten']), mz=myions['mz'], inent=myions['inten']))
        ions['bin'] = pd.cut(ions.mz, binsize, labels=binsize[:-1])
    except KeyError:
        filestem = Path(hdf5file).stem
        print(f'issue for {filestem}')
        ions = None
    return ions
```

Figure (11): Code snippets illustrating the ease with which the .hdf5 file can be used for data interrogation. With one call using Pandas (upper panel) or directly with PyTables (lower panel), all fragment ions for one spectrum are read from the .hdf5 file. Then the normalized intensity and $\log_{10}$ intensity are calculated, and the *m/z* values are binned and returned to the function caller.

the disks and the speed of the network between them, but goes some way to illustrate that gathering such information becomes a relatively trivial task and could be performed in any lab without the need for a specialized database, software or hardware configurations. This feature of isobarQuant is further showcased in Part IV of this report where it was used to profile the effect of TMT reporter group on the fragmentation of peptides using HCD.

3.2 Comparison between isobarQuant and MaxQuant

At the time of the development and later publication of isobarQuant, the popular software MaxQuant was unable to perform any isobaric quantification and, at the time, no other MS software tool was able to measure and correct for ratio compression. However, in order to make this report as complete as possible, a comparison was made between MaxQuant and isobarQuant using a serial dilution of *E. coli* peptides at known concentrations, labeled with TMT10 and spiked into a background of digested protein from a standard human cell line present at approximately equal



(a) 1% FDR-filtered *E. coli* peptides identified by MaxQuant (blue) or isobarQuant (red)

(b) 1% FDR-filtered human peptides identified by MaxQuant (blue) or isobarQuant (red)

Figure (12): Counts of 1% FDR-filtered *E. coli* and human peptides identified in the same single 115 minute run by either MaxQuant (Andromeda) or isobarQuant (Mascot). The number of *E. coli* peptides is about one third the total human peptide count. All decoy and contaminant peptides were discarded and peptides mapping solely to *E. coli* or human were tallied. Both softwares return a high number of shared peptides, with Mascot and isobarQuant yielding approximately 10% more.

amounts, or completely absent, and labeled with TMT10 reagent see table 2 in Methods above (2.4). This should allow us to make a direct comparison between samples with a high background (and associated ratio compression) and those with none (or very low) and how well isobarQuant performs to correct for this.

3.2.1 Comparison using *E. coli* dilution series spiked into human background

isobarQuant and MaxQuant were run as described above in the methods section. Retaining only peptides passing the 1% FDR filter, the numbers of identified peptides shared between Mascot and Andromeda are comparable. For both softwares, the number of *E. coli* peptides is approximately one third the number mapped to human proteins. Mascot and isobarQuant match around 10% more peptides overall than MaxQuant and Andromeda (Fig. 12).

The accuracy and precision of quantification of isobarQuant is higher than MaxQuant (Fig. 13). In all cases, the median of all fold changes estimated by isobarQuant are closer to the expected value that the median calculated by MaxQuant, when data are filtered using the default values. Here a PIF of >0.75 was used for MaxQuant (resulting in a total of 4473 unique, quantifiable spectra), compared to more stringent filter criteria employed by isobarQuant [S2I > 0.75, P2T>4, Mascot score >15 and minimum peptide length of 7 amino acids] to give just under half the number of quantifiable spectra (2208). This fact still holds when MaxQuant peptides are filtered by the more stringent isobarQuant criteria. In figure 13a, the effect of the presence of the human background on the *E. coli* ratios is clear. In all cases, and with both softwares, the median fold changes are off by an average of 0.13 for isobarQuant and 0.25 for MaxQuant. When MaxQuant peptides are filtered by the same criteria as isobarQuant, this figure drops to 0.18 (fig. 13b). For samples with no human background both softwares perform similarly well. Here the median fold change for isobarQuant is off by 0.05 when averaged over all samples with no human background, while MaxQuant is off by only 0.06 (with no change when filtering by the same parameters as isobarQuant). It is interesting to note that both softwares underestimate the median fold change in all cases where no human background was present.

3.3 Ratio-compression mitigation for published dataset

Following the download of 30 raw files isobarQuant was used to process the same raw files twice, the first time, as per default, with ratio compression correction switched on and a second time with it switched off (it is possible to perform this on the same .hdf5 file containing the internalized Mascot results and quantification data). The distribution of the fold changes for each pipetted ratio were visualized in a box plot, arranged in order of spiked in amount. In total, 94 out of the 97 spiked-in peptides were identified by a total of 4231 PSMs. Of these PSMs, 3320 were quantified and had an S2I value > 0.5 and 1988 (46% of the identified total) passed all isobarQuant quantification criteria, such as <1% FDR, minimum Mascot score of 15, minimum peptide length of eight residues, P2T > 4 and for this assessment had an S2I of > 0.75. In all cases, the median of the ratio-corrected values was closest to the pipetted, spiked-in value when ratio compression correction was performed (Fig. 14) and for all but one reporter ion channel (the 1:1 sample) the closest value to the expected ratio was in the 'high quality' filtered sample. The best improvement (compared to non-corrected, loosely filtered peptides, gray boxes, figure 14) was for the 1:10 spiked in corrected sample where the median fold-change increased from 2.24 to 6.72. With increasing ratio (higher amount of spiked-in peptide) the IQR also gets wider. This phenomenon seems to be amplified following data filtering and also after ratio compression correction, the IQR increasing from 6.29 to 7.78 in the corrected, 1:10 pipetted sample and from 2.24 to 5.03 in the uncorrected sample. The IQRs of the data from the smaller fold change samples was much lower and much more comparable between corrected and non-corrected, filtered and non filtered data.

3.4 R-based graphical outputs

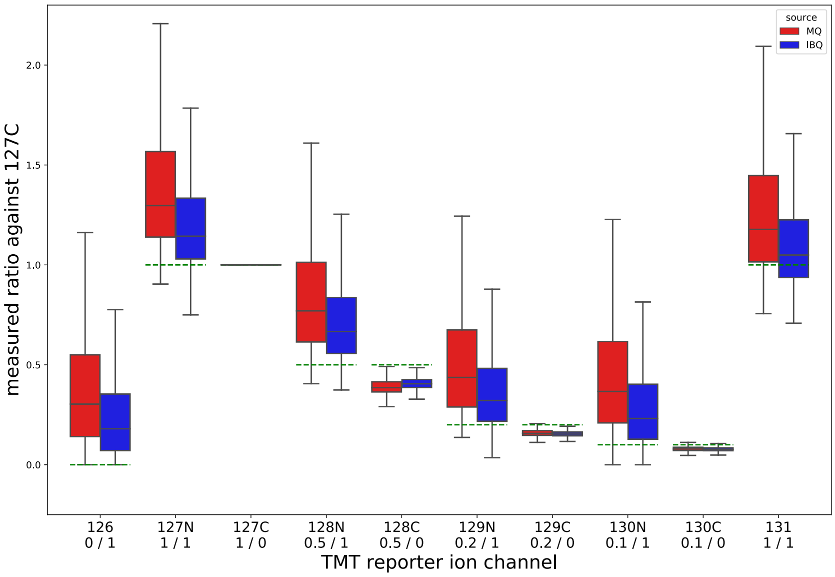
In addition to the above outputs an R script giving summary information for the files acquired during the whole run is a useful addition. This is part of the isobarQuant downloadable package and provides metrics about the performance of the instrumentation for each individual run. There are twelve different plots which are summarized in the table below. Example plots may be found in the appendix.

Table (4): Overview of R script output for instrument QC purposes

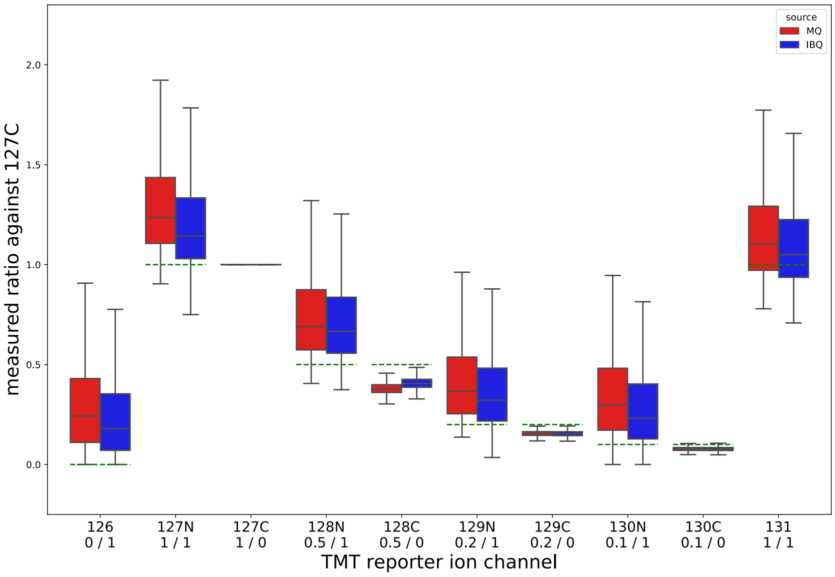
page in PDF output	Output	Description
1	Acquisition overview table	Displays time, date and length of acquisition plus instrument name

Table (4): Overview of R script output for instrument QC purposes

page in PDF output	Output	Description
1	Search results table	Displays Mascot search result information such numbers of MS/MS assigned (unique / all) , number of hook peptides and mean ppm error for precursors
1	Search details table	Shows search database used with precursor and fragment ion tolerance used
2	Log10 TIC precursor	Distribution of $\log_{10}$ transformed precursor ion intensities
2	Log10 TIC MS/MS	Distribution of $\log_{10}$ transformed total ion current intensities for all fragment ions
2	Log10 TIC MS vs MS/MS	Log_{10}transformed precursor intensities as function of $\log_{10}$ transformed TIC intensities of MS/MS spectrum
3	Log10 precursor over precursor *m/z*	Scatter plot of $\log_{10}$ transformed precursor intensities for precursor *m/z*'s
3	Log10 MS/MS over precursor *m/z*	Scatter plot of $\log_{10}$ transformed TIC of MS/MS fragment ion intensities plotted per precursor *m/z*'s
3	Precursor ppm delta over *m/z*	Scatter plot of deviation of precursor *m/z* from expected values
4	Distribution of precursor charge states	Histogram of different precursor charge states recorded throughout run.
4	Distribution of MS/MS events per precursor	Number of MS/MS events triggered from a single MS peak selected for fragmentation
5	Basepeak chromatogram	Basepeak chromatogram per minute retention time (RT)of the run
5	Frequency and fate of MS/MS events	Histogram of frequency of MS/MS events per minute RT bin with the corresponding success rate encoded by color. Success is measured as no match found by search engine, match to a PSM and match to high confidence (hook) PSM
6	cycle time between MS1 scans	Density plot of cycle time between consecutive MS1 Scans
7	MS/MS rap fill time	Density plot of time taken to fill Orbitrap for MS/MS spectra
8	Mascot Score Distributions	Mascot score distributions for all rank 1 PSMs shown as a violin plot overlaid by a box plot, where the interquartile range is represented by the upper and lower box edges. The median score value and total number of PSM matches (n) is given
9	Mascot Score distributions by decoy / target	Mascot score distributions for all rank 1 PSMs as above but divided into target and decoy peptides. Display is a violin plot overlaid by a box plot, where the interquartile range is represented by the upper and lower box edges. The median score value and total number of PSM matches (n) is given
10	True spectra vs FDR	receiver operating characteristic (ROC) curve of cumulative true positive PSMs at each estimated FDR
11	FDR vs Mascot score	Plot of cumulative FDR for Mascot scores of PSMs to yield the corresponding q-value. The q-value equaling an FDR of 1
12	Distribution of P2T	Violin and box plot of P2T values
13	Distribution of S2I	Violin and box plot of S2I values
14	Distribution of FWHM	Distribution of FWHM of chromatographic peaks
15	Distribution of distance MS/MS to Apex	Distribution of time difference between apex of XIC and RT of triggered MS/MS event
16	Distribution of noise cut offs at *m/z* 128	Distribution of intensities of instrument noise around *m/z* 128 for MS and MS/MS over retention times of run
16	Distribution of noise cut offs at *m/z* 500	Distribution of intensities of instrument noise around *m/z* 500 for MS and MS/MS over retention times of run
16	Distribution of noise cut offs at *m/z* 800	Distribution of intensities of instrument noise around *m/z* 800 for MS and MS/MS over retention times of run



(a) All peptides quantified and filtered by MaxQuant (red) [4473] and isobarQuant (blue) [2208]. Peptide ratios estimated by isobarQuant are closer to the pipetted values.



(b) The 2208 *E. coli* peptides shared between MaxQuant (red) and isobarQuant (blue). Filtering of MaxQuant peptides by isobarQuant criteria brings the median of the fold change ratios closer into line with the expected values and the IQR of the distributions are narrowed. Center line in box plots is the median, the bounds of the boxes are the 75 and 25% percentiles i.e., the IQR and the whiskers correspond to the highest or lowest respective value or if the lowest or highest value is an outlier (greater than 1.5 * IQR from the bounds of the boxes) it is exactly 1.5 * IQR

Figure (13): Box plot of peptide fold changes for *E. coli* peptides following processing of an identical .raw file with MaxQuant and isobarQuant. *E. coli* samples were labeled in a dilution series and half were added to a human TMT-labeled sample. All human protein was present at the same amount. The pipetted ratio for *E. coli* is shown on the lower *x*-axis label before the slash and the pipetted human ratio is given after it. The pipetted (expected) values shown as dashed green line. Fold changes were calculated by dividing all reporter ion areas by the baseline *E. coli* signal with no human background. Fold change ratios with no human background show much higher accuracy and greater precision, though in all cases are over estimated. . The box plots in the upper panel are based on the default settings for quantification of peptides in either software. In the lower panel, the box plots highlight the differences in fold change between peptides from both softwares filtered according to isobarQuant filter criteria.

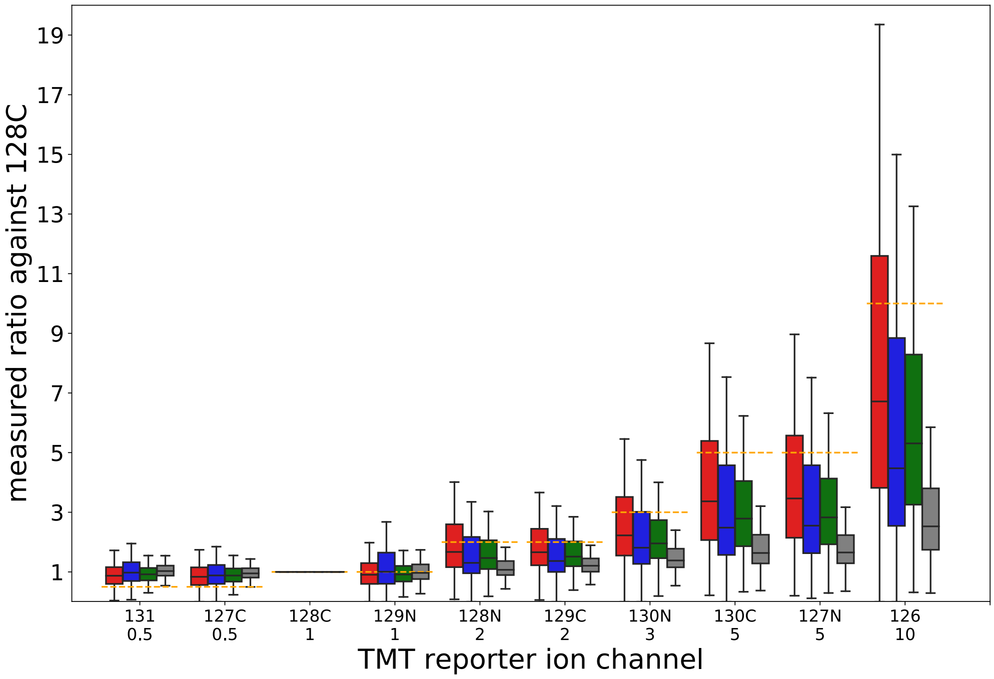


Figure (14): Box plots to highlight the effect of ratio compression correction in isobarQuant for the Keshishian dataset. Red boxes indicate the distribution of measured TMT fold changes after correction by isobarQuant and following filtering to include only peptides with a recorded S2I value of 0.75 or higher and fulfilling all other default isobarQuant quantification criteria. This 'high qua1ity' set consists of 1988 PSMs mapping to 94 non-redundant peptides. The data represented in the blue box plots are corrected by isobarQuant but not part of the high-quality set. These are only filtered for an S2I of greater than 0.5 and yield 1232 quantified PSMs mapping to 94 non-redundant peptides. Green box plots are data which are not corrected by isobarQuant for the effects of ratio-compression but are filtered by the same criteria high-quality criteria. The gray boxes display the ratios of the non-corrected PSMs, filtered only to retain all those with an S2I greater than 0.5 (as with the blue box plots). The expected (or pipetted) values are indicated by the dashed, orange line. The *x*-axis is marked with the TMT reporter ion channel and under it the spiked-in ratio with channel 128C as the denominator. The center line in the box plots is the median, the bounds of the boxes are the 75 and 25% percentiles i.e., the IQR and the whiskers correspond to the highest or lowest respective value or if the lowest or highest value is an outlier (greater than 1.5 * IQR from the bounds of the boxes) it is exactly 1.5 * IQR

4 Discussion

With the release of isobarQuant the proteomics community was given access to state-of-the art methods developed in order to computationally analyze the results of isobaric quantification in mass spectrometry experiments, with a particular emphasis on off-target proteomic analyses, which at the time of development was not covered by any other tools. At the time of writing it is (knowingly) being used in more than five labs across the world and its use is referenced by more than twenty publications[178–202], many of which perform the off-target proteomic analyses made possible through the TPP experiments.

Its text-based outputs are simple and can be visualized without any additional software or tools. They contain only the data that is most relevant or required for use with downstream analysis tools. However, all experimental and a great deal of 'meta' data continues to be available within easy reach inside the .hdf5 files. Among the papers cited to use isobarQuant one stated that it only used the peptides.txt output of isobarQuant in a non-quantitative experiment: so the simplicity of the output appeals not only to experimenters performing quantitative experiments. With a minimal amount of customization, isobarQuant was also successfully used to internalize and re-analyze a very large data set that was no longer possible to view on a Mascot server, probably because the large size of the dataset rendered this impossible.

4.1 Comparison to MaxQuant

The presence of human background in an *E. coli* dilution series leads to systematic compression of *E. coli* signals in the direction of ratio 1:1. The situation is improved by processing data with isobarQuant compared to processing with MaxQuant but is still not completely resolved. Both softwares perform equally when estimating fold changes in the presence of no or low background. In this experiment a small underestimation of the fold change ratios occurs in the background-free experiments, but this is likely caused by increased signal in the baseline sample, probably due to slight over-correction of reporter ions during isotope impurity correction or possibly a small amount of ion coalescence from neighboring reporter ions during acquisition. The apparent compression of the fold change ratio in the 126 labeled sample is surprisingly high, since no signal from *E. coli* was expected in that channel. Fold changes determined by isobarQuant show increased accuracy over those determined by MaxQuant for an identical set of peptides, demonstrating that the increased accuracy was not solely brought about by the exclusion of peptides with lower fold change accuracies, but by the ratio-compression (S2I) correction. The numbers of peptides identified by MaxQuant (Andromeda) and isobarQuant (Mascot) showed a good overlap with approximately 10% more identified by Mascot at 1% FDR. In summary, isobarQuant offers a larger set of filter criteria to exclude peptides with less accurate quantification signals than MaxQuant and is able to correct for ratio compression which is not currently possible within MaxQuant.

4.2 Comparison to published dataset

The fold changes calculated by isobarQuant were more accurate than those published by Keshishian and colleagues[177] for spiked-in peptides across different dilutions in a complex background. The improvement was brought about not only by more stringent filtering criteria but also by ratio compression correction. In all cases, the median peptide fold changes were shifted closer to their expected (pipetted) values. With increasing pipetted ratios, in-sample variability also increases. This apparent drop in precision seems to be amplified upon both peptide filtering and ratio-compression correction at higher fold change ratios. It was not removed by filtering out peptides with overall lower reporter ion intensities and is likely a result of experimental 'wobble' associated with decreased pipetting accuracy at higher ratios. It cannot be attributed to ratio compression correction since the decrease in precision is also seen in the non-corrected, loosely-filtered data and appears to be inversely proportional to higher expected ratios. Overall an improvement is brought about compared to the fold changes originally calculated with results closer to the expected values resulting from the improved filtering and ratio compression correction.

4.3 General remarks

To the best of the author's knowledge isobarQuant was the first freely-available, proteomics software tool that allowed the experimentalists to computationally correct for ratio compression. Its simplicity of use was further demonstrated when the isobaric TMT tag was extended to fill all possible 11 channels in 2017. The only addition needed for isobarQuant was the additional mass for the 'heavy' 131 reporter ion fragment at *m/z* 131.144999 (more accurately termed the '131C' ion) along with the corresponding additional correction factors to perform the heavy isotope impurity correction); the identical situation occurred a second time in 2019 with the announcement of new isobaric reagent, TMTpro, a sixteen-plex extended variant of TMT. isobarQuant worked as soon as the new mass values were added to the configuration and it was immediately put to use in analyzing the new quantification tag. isobarQuant's construction is generic enough that the new outputs simply had one (or six) extra columns.

The performance of isobarQuant was dramatically improved by adding the multithreading capability bringing the total processing time for 12 .raw files (the typical number for a TPP single concentration offline fractionated experiment) down to 10 hours from start to finish on a standard laptop. The pivotal role of the .hdf5 file and PyTables in the speed and performance of isobarQuant cannot be underestimated. It was an essential part of the investigation into peptide fragmentation in part IV and enables the extraction of fragmentation data from many hundreds to thousands of spectra stored throughout many files over a large SATA file system to amount to only a few minutes. This under-explored functionality could really benefit scientists who would like to create a spectral library or build transition lists for a target proteomics approach but do not have the luxury of a large database to store the results of previous experiments all the way down to the level of fragment ions from MS/MS spectra.

5 Outlook

isobarQuant in its current format will continue to perform excellently for what it was designed to do: isobaric quantification (and also MS1-based quantification; see following chapter III). It will continue to be able to process MS1, MS2, MS^n spectra and derive protein fold changes from filtered isobaric and reporter ions. However in the years since its first publication, a couple of novel quantification methods have become more established and could warrant investment to support them in isobarQuant. For example, Neucode[203,204] which takes advantage of the mass defect between nitrogen and carbon atoms to open up a new mode multiplexed mode of MS1-based quantification using high resolution MS1 scans or the idea of using the series of ions complementary to the reporter ions for quantification[205,206]or similarly EASI-tag[69].

Also, since the first release of isobarQuant a couple of new methods have been published to reduce the effect of ratio compression[207,208], how these compare to the approach taken by isobarQuant would need to be investigated and the gain in value of incorporating the methods would need to be assessed.

There are still a few specific sections of the code that could benefit from some detailed profiling, such as the .mgf creation part which takes between five and ten minutes to create a single .mgf file, but since the overall time spent on data processing is still relatively low compared to the acquisition of the data this is not really a grave issue; secondly, as is so often the case with modern software development, it might actually be more cost-effective to throw more hardware at the problem and simply 'buy a bigger server' and run isobarQuant with more CPUs rather than spend time refactoring code.

An evaluation of isobarQuant from a pure software development side might conclude that a major reformat of isobarQuant to make it more modular would be beneficial. This would make the functionality of key aspects of the software available to any other application and could then be implemented in the other ways such as in the processing of data from different types of mass spectrometer. This would also make it easier to use isobarQuant in concert with other computational mass spectrometry software such as Percolator[124], Ursgal[127] or pyQms[209]. There would also be benefit in the integration of isobarQuant with other search engines apart from Mascot. Since its development there has been an increase in the number of published search engines (or improvements to their underlying algorithms) for example MSFragger[92], peppy[78] or MS-GF+[210] as well as alternative methods of sequence matching involving *de novo* or AI approaches[107,108]) that might offer some advantage over Mascot in certain scenarios. The first requirement here would be to at least include all Mascot peptide ranks.

The focus of the majority of the studies for which isobarQuant has been used was changes in abundance between specific gene products under differential conditions (various temperatures, different compound concentrations or other alternative settings) where it was not necessary to extend the level of granularity to beyond the gene level. Performing experiments to dissect and investigate changes at the level of alternative splice isoforms is currently possible with isobarQuant, but to really delve into and make sense of peptides in this different context will require a fresh look at the protein inference part of the software, even if the peptide and quantification parts could in essence remain as they are now.

A more pressing need to be addressed will be the shift in Python version. The Python Software Foundation will cease to support Python 2x in January 2020[211] and after this point the Python community will only continue development of Python 3x. This will not mean that isobarQuant will no longer work but, as with a large number of other systems running legacy code throughout the world, staying on Python 2x will mean that any new features brought out in Python 3x will not be accessible. The switch to Python 3x is not trivial and would require a new version of the compiled library against the Thermo Fisher Scientific DLL. It could also be the trigger to start looking into alternative ways of the accessing raw data in ways similar to mzMine[212] or the Thermo Fisher Scientific API for Linux or potentially using a Linux machine running WINE (a free and open-source compatibility layer that aims to allow computer programs developed for Microsoft Windows to run on Unix-like operating systems) as described at the UPWR (https://proteomicsresource.washington.edu/protocols06/wine/). A less convenient approach would be to first convert .raw files to mzML[213] using a software such as ThermoRawParser[214] or the MSConvert utility of ProteomeWizard[215] and enable isobarQuant to use that as its starting point. However, this has the primary disadvantage of adding quite some processing time to the pre-Mascot workflow, since files need to be converted twice, once to mzML and then a second time to .hdf5 format.The main advantage of this would be to bring isobarQuant much more in line with the current community standards such as TPP[141], openMS[216] and Galaxy[217] which use mzML (or .mgf) as their input formats.

In this author's opinion the proteomics community is currently in the first quartile of the development of quantitative proteomics. The opportunities offered by DIA; targeted methods such as SRM / MRM and PRM; MS1 precursor *glsm/z* window selection techniques such as Boxcar[218]; targeted quantitative approaches like TOMAHAQ[37] and of course the continued advances in label free quantification through ever improved methods. isobarQuant is well placed to be part of it.

6 Author Contribution to project

The results of this project were published in ref.[161], as part of the isobarQuant package, a constituent of the suite of software for processing experiments in Thermal Proteome Profiling mode.

Toby Mathieson co-designed the layout and initial implementation of isobarQuant and wrote the post-Mascot part of the pipeline, contributing significantly to the pre-Mascot part of isobarQuant in terms of design, concept and coding. All data analyses presented in this section were conceptualized and performed by the author. The author implemented the multi-threading capability of both post- and pre-Mascot workflows and designed and coded all Mascot-server specific parts of the pipeline. All experiments to assess and compare the results of mitigation of ratio compression to the MaxQuant software was conceptualized by the author and the laboratory experiments were performed at Cellzome, a GSK company as part of a standard workflow QC step. Since its publication, the author has been the first point of contact for all external communications related to the project with several different international laboratories.

Part III

Systematic analysis of protein turnover in primary cells - an extension to isobarQuant to allow peptide ion (MS1)-based quantification

1 Introduction

In its first implementation isobarQuant focused on the determination of relative protein abundance via fragment ion intensities from isobaric tags. This technique is well suited to differential quantification experiments where changes between different conditions are sought, but it is not as accurate as peptide ion (MS1) intensity-based methods. In the next project presented in this PhD thesis the intention was to use SILAC to measure global protein turnover is several primary cell types in a pulsed (dynamic) workflow. Here, precise and accurate MS1 quantification is paramount, since even small deviations in the accuracy of measured fold changes can have a pronounced effect on the half-life measurement[219]. This is particularly acute when measuring protein turnover in non-dividing cells[220] since many proteins exhibit a very slow turnover because the replication of the entire proteome, which normally occurs in exponentially growing cells, is not taking place. In these cases missing quantification values (the result of peptide ion signals dropping below the instrument-determined noise thresholds[173]) can also become problematic. Secondly, primary cells can only be kept in culture for a limited amount of time before adapting to the cell culture conditions or going into senescence, hence protein turnover determinations must be based on relatively short-term treatments with stable isotope-encoded amino acids.

To enable reliable half-life determination from these very small changes in protein abundance the author developed methods based on a better utilization of the isotopic distributions of ionized peptides and their associated features and incorporated these into isobarQuant. A great deal of time was spent in optimizing the algorithms to ensure the highest quantified protein coverage but using only the most accurate peptide fold changes. The resulting improvements in isobarQuant led to the publication of a catalog of more than 9, 600 protein half-lives across five different primary cell types.

1.1 MS1-based quantification in other software

At the time of development of this part of isobarQuant there was already some software that could perform MS1-based quantification available: MaxQuant[138], PyQuant[221] and ProteomeDiscoverer (Thermo Fisher Scientific, San Jose, CA, USA[144]) but their focus had primarily been on higher ratio fold changes after label incorporation in growing cells and the precision and accuracy for quantification of very low protein ratios had not been presented. Additionally, with the simple isobarQuant framework already in place, it would make sense to be able to process data using isobarQuant and to create outputs in a familiar, similar layout; not only for the ease of experimenters but also for any downstream tools that might be reliant on a given structure.

Experiments would be performed in five different primary cell types. Four from human: B-cells; natural killer (NK) cells; undifferentiated monocytes; and one non-immune cell type hepatocytes; and from mouse came embryonic neurons. In these non-dividing cells, the incorporation of heavy isotope labels will be very slow for some proteins and will consequently be error prone particularly at the early time points. Understandably this demands high confidence in the measurements we make but equally this should not come at the cost of low coverage. The goal was therefore to investigate and optimize the parameters leading to protein fold changes which are as accurate, precise and reproducible as possible and also for the highest obtainable number of proteins.

Most MS1 matching software uses an averagine model fitted to an XIC of acquired values. The averagine model, making use of a virtual amino acid 'averagine', constructed using the statistical occurrences of amino acids in the human proteome, is often used to estimate a precursor's isotopic envelope because the elemental composition of the precursor is not known until after the peptide identification step is performed at a later stage in the workflow[222]. Consequently, any measure of fit between the intensities will be inaccurate compared to a model that is based on the actual peptide being identified. This limitation is not likely to have a strong impact when dealing with large fold changes but gains in importance when the intensity ratio is very low. Equally important at these very low ratios is a measure to detect overlapping isotopic distributions from different peptides which could mask the true signal. The difference of using an exact model versus an averagine model is illustrated for a theoretical peptide in figure 1.

2 Methods and implementation

In order to determine half-lives of proteins in five different types of non-dividing cells, a pulsed-SILAC approach was applied. The mass spectrometry data was acquired on a high resolution Thermo Fisher Scientific Q-Exactive instrument. In order to obtain the most accurate and precise data, the isobarQuant software package was adapted and used to process the acquired data.

2.1 Preparation of samples

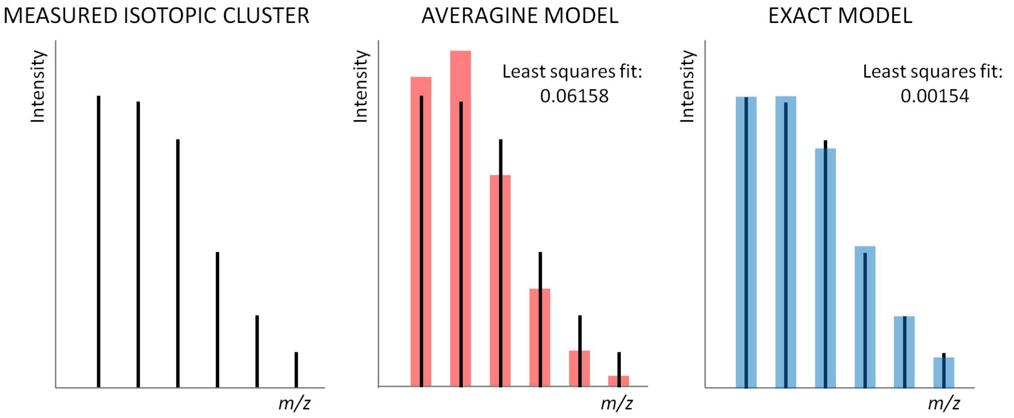


Figure (1): Example of fitting two different intensity models to a measured isotopic cluster. The averagine model, here depicted by red bars is fitted to the acquired intensities with a worse least squares fit than the exact model, here shown as blue bars. Following identification of a peptide and modification(s) by the Mascot search engine it is possible to construct an exact model from its elemental composition. An intensity for the heavy or light peptide is calculated by fitting the corresponding exact model of the peptide (here the SILAC-light form of peptide GPCMSE-QAMGPCMSEQAMK) to the acquired data using a least-squares method.

Time point	B-cell replicates	NK cell replicates	Monocytes replicates	Hepatocytes-replicates	Mouse Neurons
1	7 hours	7 hours	7 hours	9 hours	6 hours
2	11 hours	11 hours	12 hours	12 hours	12 hours
3	24 hours	25 hours	24 hours	27 hours	24 hours
4	34 hours	35 hours	36 hours	75 hours	35 hours*

* second replicate: 36 hours

Table (5): Times of pulse-in experiments

2.1.0.1 Primary cell isolation and treatment Primary human hepatocytes (KaLy Cell) and human monocytes, B-cells and NK cells, isolated from peripheral blood mononuclear cell (PBMC)s derived from buffy coats (German Red Cross, Mannheim) by magnetic-bead based negative selection (STEMCELL Technologies), were adapted to the light (L) SILAC medium overnight at 37°C. Cells were then pulse-labeled with heavy (H) isotope-labeled amino acids (lysine, ($^{13}C_6{}^{15}N_2$, Sigma-Aldrich, 608041) and arginine ($^{13}C_6{}^{15}N_4$, Thermo Fisher Scientific, 88434)) for the indicated time periods: Table (5), washed, pelleted, and snap-frozen in liquid N_2. Cell pellets were lysed in buffer containing 4% SDS and digested with benzonase.

2.1.0.2 Primary neuron culture Cortical neuronal cells were isolated from prenatal embryos of CD-1 mouse at embryonic day 15 (E15). To dissociate the cortex tissue, it was finely chopped by scalpel followed by digestion in Accutase (ThermoFisher, A1110501) for 12 mins. To prevent clumping due to DNA from dead cells, tissue was treated with 250 unit/µl of benzonase (Millipore, 71206-3). Neurons were triturated gently with a fire-polished Pasteur pipette and passed through the 40 µm cell strainer (BD Falcon, 352340) before plating them onto 6 well plate at a density of 1x10^6 cells per well. The plates were coated with 0.1 mg/ml of poly-D-lysine (Sigma-Aldrich, P0899) and 2.5 µg/ml of laminin (Sigma-Aldrich, 11243217001). Cultures were maintained in Neurobasal medium (Thermo Fisher Scientific, 21103) containing 1% penicillin / streptomycin (ThermoFisher, 15140122), 1% GlutaMAX (ThermoFisher, 35050), and 2% B27 supplement (ThermoFisher, 12587) at 37°C with 5% carbon dioxide in the incubator.

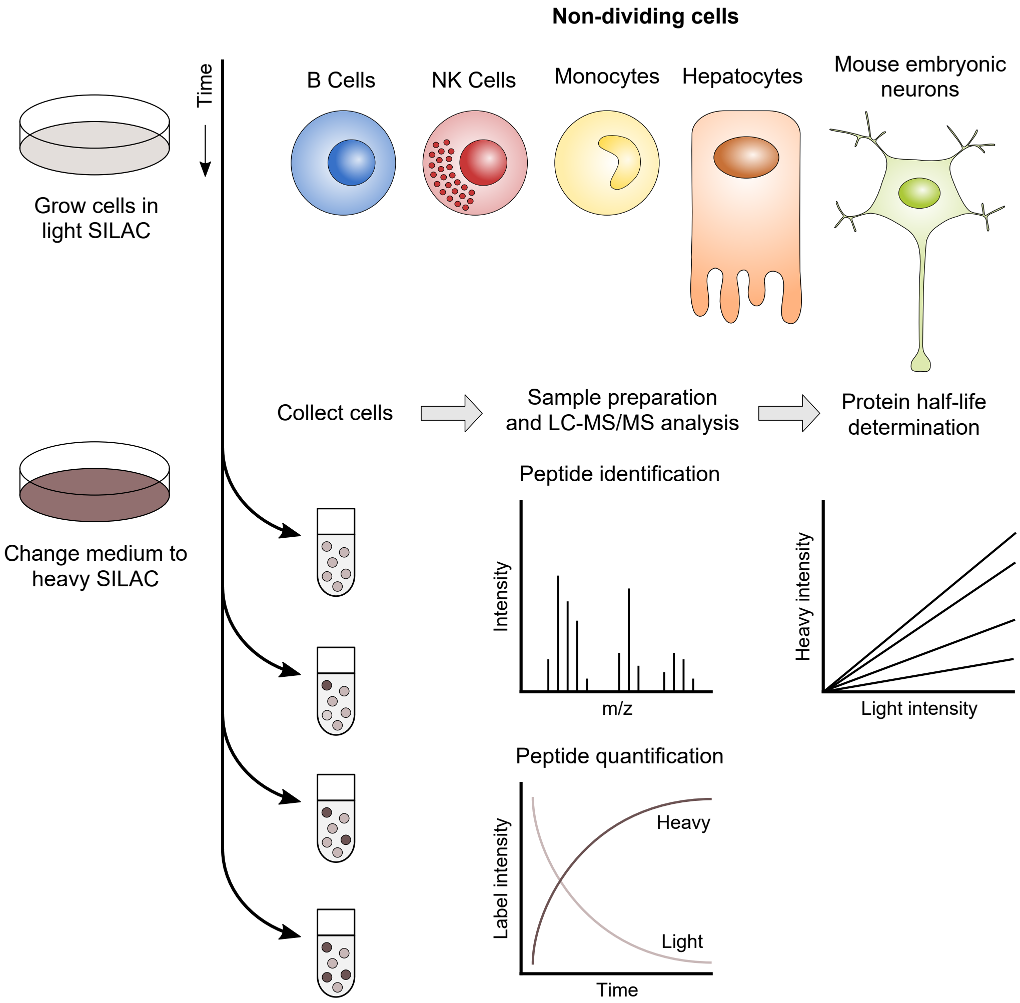


Figure (2): Workflow for determination of protein half-lives taken from ref.[181]. To label newly synthesized proteins, the cells were exposed to heavy SILAC medium and collected at different time points. After protein extraction, proteolysis with trypsin, sample preparation, and subsequent LC-MS/MS analysis, the peptides were identified by the Mascot search engine and quantified via isobarQuant. Peptides of pre-existing and newly synthesized proteins were distinguished by their mass due to incorporation of light or heavy arginine and lysine. Protein fold changes at different time points were calculated using the intensity ratios of heavy vs. light SILAC peptides and were used for subsequent protein half-life determination.

Post-seeding after 1 day *in vitro* (DIV 1), half of the medium was replaced with fresh pre-warmed Neurobasal medium with all the supplements (above). On DIV 4, neurons were treated with 1.0 µM of cytosine arabinoside (Tocris, 4520/50). On DIV 5 dynamic SILAC experiments were started by exchanging one fifth of the medium with a final 10x excess of heavy lysine ($^{13}C_6{}^{15}N_2$, Sigma-Aldrich, 608041) and heavy arginine ($^{13}C_6{}^{15}N_4$, ThermoFisher, 88434). Cells were harvested 0 hours, 6 hours, 12 hours, 24 hours or 36 hours after pulse, washed with PBS including protease inhibitors (Sigma-Aldrich, CO-RO Roche) and lysed in 50mM Tris-HCl, pH7.4, supplemented with 4% SDS and benzonase. Lysates were cleared by centrifugation at 20,000 xg at room temperature, followed by protein concentration measurement (BCA assay, Thermo Fisher Scientific, 23225). 20µg protein of each time point were used for MS analysis.

All animals were housed in the EMBL animal facilities under veterinarian supervision and are treated following the guidelines of the European Commission, revised directive 2010/63/EU and AVMA guidelines 2007.

2.1.0.3 THP-1 SILAC cell mixtures for method evaluation THP-1 cell cultures (ATCC TIB-202) were established in RPMI-based SILAC media and supplemented with either light or heavy isotope labeled-amino acids (as above, 2.1.0.2). For harvesting, the cells were washed, pelleted, and snap-frozen in liquid N_2. They were lysed with 4% SDS in 50 mM Tris-HCl pH 7.4 and the DNA was digested with benzonase nuclease (Sigma, E1014-25KU). Three independent dilution series (1:1, 1:9 and 1:49) were created by mixing the 25 µl, 5 µl and 1 µl (respectively) of SILAC-H to SILAC-L medium to a final volume of 50µl.

2.1.0.4 Sample preparation for mass spectrometry Cells were washed with PBS and the supernatant was removed completely before cells were lysed in 2 % SDS for 3 min at 95 °C in a thermomixer (Thermo Fisher Scientific), followed by digestion of DNA with benzonase at 37 °C for 1.5 hours. Lysate was cleared by centrifugation and the protein concentration in the supernatant was determined by BCA assay. Proteins were reduced by DTT and alkylated with iodacetamide, separated on 4–12% NuPAGE (Invitrogen), and stained with colloidal co-omassie[223] before proceeding to trypsin digestion and mass spectrometry analysis (see 2.1.0.5). Gel lanes were cut into three slices covering the entire separation range (~2 cm) and subjected to in-gel digestion[224]. Peptide extracts were additionally fractionated on an Ultimate3000 (Dionex, Sunnyvale, CA) using reversed-phase chromatography at pH 12 [1 mm Xbridge column (Waters, Milford, MA)], as described in ref.[225].

2.1.0.5 LC-MS/MS analysis Samples were dried *in vacuo* and re-suspended in 0.05 % TFA in water. Of the sample, 50% was injected into an Ultimate3000 nanoRLSC (Dionex, Sunnyvale, CA) coupled to a Q-Exactive plus (Thermo Fisher Scientific). Peptides were trapped on a 5mm x 300 µm C18 column (Pepmap100, 5 µm, 300 Å, Thermo Fisher Scientific) in water with 0.05 % TFA at 60 °C. Separation was performed on custom 50 cm × 100 µM (ID) reversed-phase columns (Reprosil) at 55°C. Gradient elution was performed from 2% acetonitrile to 40% acetonitrile in 0.1% formic acid and 3.5% DMSO over two hours. Samples were online injected into Q-Exactive plus mass spectrometers operating with a data-dependent top 10 method. MS spectra were acquired using 70,000 resolution and an ion target of 3x10^6. HCD scans were performed with 25% nCE at 17 500 resolution (at *m/z* 200), and the ion target setting was fixed at 1x10^6. The instruments were operated with Tune 2.3 and Xcalibur 3.0.63.

2.2 Post acquisition analysis

The isobarQuant suite of software was adapted to perform peptide precursor intensity based quantification by including a module that is triggered after the Mascot parser step is finished. It is described below. The code was adapted so that the methods associated with isobaric quantification were not invoked. The switch between quantification modes is determined by the quantification method supplied on the command line at the start of the pre-Mascot. workflow. The masses related to SILAC quantification were added to the configuration file (QuantMethod.cfg) in the section ['silac3'] (as it is possible to perform SILAC in three modes, although for this experiment only two were used, heavy and light): for LIGHT (K+0, R+0), MEDIUM (K+13C6, R+13C6) and HEAVY (K+13C6+15N2, R+13C6+15N4) SILAC modifications on lysine and arginine. The quantification source was set to MS1.

2.2.0.1 Pre-Mascot workflow Data were processed using the pre-Mascot workflow as described in part II and then searched with Mascot 2.5 via the Mascot Daemon version 2.5.1 against the October 2014 release of Human Uniprot Proteome appended with a reverse decoy version of the same with 10 ppm mass tolerance for peptide precursors and a 20mDa tolerance for fragment ions (since high resolution data were acquired in HCD mode in the Orbitrap). Carbamidomethylation of cysteine residues was selected as a fixed modification and the following were selected as variable modifications: oxidation of methionine, acetylation of protein N-termini, SILAC heavy label 13C(6) 15N(4) on arginine (+10.008269 Da), and SILAC heavy label 13C(6) 15N(2) on lysine (+8.014199Da).

2.2.0.2 Quantification of peptides and calculation of intensity fits This was the point within the existing workflow that the first adaptations were made to include MS1 based quantification. The post-Mascot workflow of isobarQuant was started in 'mergeresults' mode to merge data acquired from the multiple offline fractionation steps. The first part of the workflow extracts the relevant data from the Mascot results files as described in 2.2.1 to ensure that the links were made between the identified peptides and their precursor ions. Following this internalization, the algorithm proceeds as follows:

1. Peptides are selected from the peptides table in the .hdf5 file and filtered to only include Mascot rank 1 peptides. Based on the peptide sequence, charge state and any modifications [not related to quantification (e.g. heavy SILAC

arginine)] all PSMs are condensed into groups, with the highest Mascot scoring PSM representing each group as a single record. This single record is referred to as a PCM. It is assumed here that the best Mascot score is a proxy for the PSM present at highest abundance. Additional information is selected from the msmsheader table in the .hdf5 file (precursor *m/z*), retention time, retention time apex, and survey scan ID)

2. The PCMs are processed in ascending order of RT. Later on this expedites the selection of relevant precursor data. Based on the PCM's monoisotopic mass, theoretical isotope mass distributions and intensities are calculated for both labels (H/L) and the charge associated with it. The theoretical masses and intensities are derived either using the exact model (based on the atomic values of the peptide and any modification) or on the averagine model using the average for atoms in 20 amino acids. The choice of model to use for quantification can be configured to use the averagine. During the development phase both values were kept and stored for use in the downstream calculations and assessments.

3. The raw XICs present within a one-minute window around the PCM's retention time (and within an 8mDa / 8ppm tolerance of recorded *m/z*) are extracted from the raw data for each isotope mass.

4. Chromatographic peaks in each XIC were detected and grouped to form isotopic clusters for each label state. The grouping involved the mapping of overlapping peaks identified from each isotope XIC to peaks identified from the monoisotopic XIC. Peaks are only considered to be overlapping when the RT spread between the two 50% apex intensity points of the peaks overlap. All possible clusters are generated before removing any where the first ^{13}C isotope is missing. If switched 'on' in the configuration, at this point the isotope clusters are also generated from the preceding survey spectrum (PS) just prior to the MS2 spectrum acquisition (of the PCM) and also from the survey scan at the apex (AS).

5. According to the given PCM an exact model of the intensities of the theoretical isotopic envelope, based on its elemental composition, was constructed. This is depicted below (Fig. 3) and is carried out for all available labels (here: the heavy and light SILAC versions of the PCM).

6. A least-squares method was used to find the isotopic cluster with the best fit to the exact model from the list of candidates in the XIC data. This yields two values: a fitted intensity used for quantification and a measure of the quality of the fit, calculated as the sum of the squares of the residual values. If the fit of the best of the XIC clusters is greater than 0.1 it is compared to the fits obtained from the AS and PS and the best fitted result of the three is selected.

7. The quantification value, reported and stored in the .hdf5 file, is the sum of the theoretical intensities calculated for the label multiplied by the fitted intensity.

2.3 Assessment of averagine and exact model

To test the assumption that a theoretical isotopic envelope based on the exact model matches more accurately to the identified isotopic peaks than the averagine model, the thirteen .raw files originating from the monocytes sample, harvested 7 hours after swapping to heavy medium were processed using isobarQuant. The peptides were filtered for Mascot score > 15, passing 1 % FDR threshold and the values representing the quality of the least squares fits of the peptides (calculated as the sum of the squares of the residual values, referred to as the least squares fits from here on) were stored in the .hdf5 file. The values for either model were -$\log_{10}$transformed and plotted for the same precursors, firstly for light and then for heavy peptides.

2.4 Calculation of prior ion ratio

The ‘prior ion’ is defined as the peak occurring at an *m/z* corresponding to the loss of one neutron from the monoisotopic ion. This peak is not expected to be present in light peptides and to have a low intensity for heavy peptides because it is usually the result of incomplete incorporation of heavy atoms into the heavy SILAC label; this is shown in left panel, Fig. 3. If an intense peak is present at this position it indicates that a co-eluting (interfering) isotope cluster is present and that the fitted intensity is likely to be overestimated (right panel, Fig. 3).

2.5 Determination of optimal settings and implementation within isobarQuant

For this part of the analysis the same monocyte data from the early time point (7 hours after swapping to heavy medium) was assessed to see the effects of different cut offs on the total number of peptides and proteins quantified using both the exact and the averagine model. The effect on reproducibility / precision was also examined. The early time point was chosen as it would represent the most challenging setting where the fold changes would still be very low. The relevant data was extracted from the .hdf5 files and processed using R or Python and plotted in R, Python or GraphPad Prism.

2.5.1 Assessment of least squares cut off

isobarQuant which was adapted as described above to store the least squares fits and calculated prior ion ratios to all PCMs (both heavy and light labels) was run once using the averagine model and once using the exact model for each of the monocyte datasets (both replicates) in ’merged results’ mode. Protein inference and FDR calculations were performed as described above (2.2.2). Having then applied the ’standard’ filters for uniqueness, Mascot score > 15, peptide length $\geq$ 6 and FDR < 1 % it was

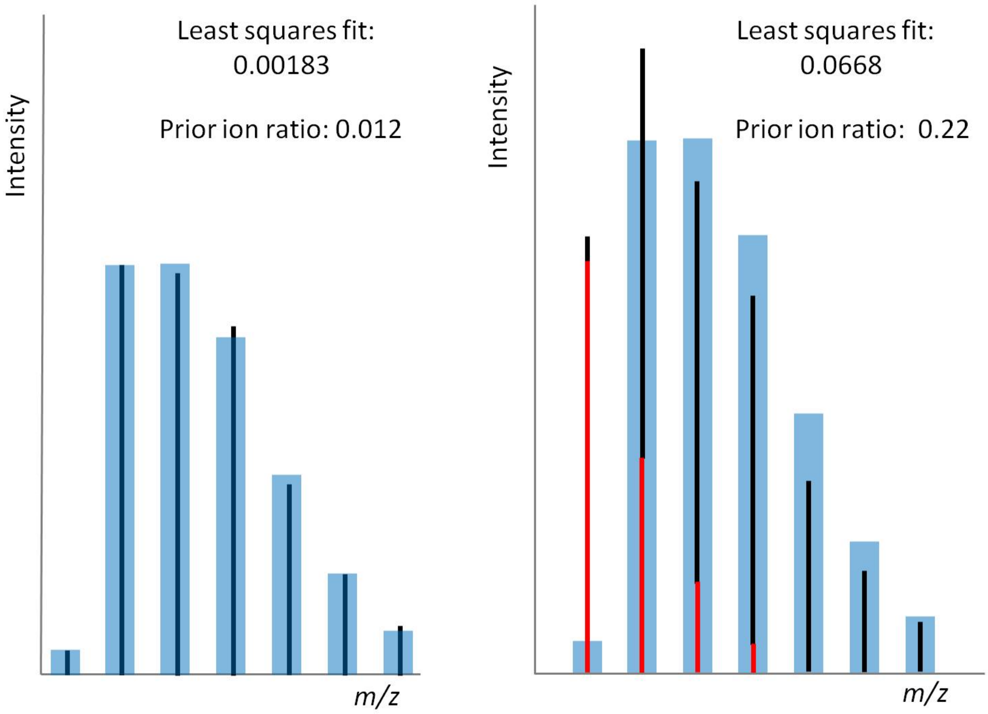


Figure (3): Depiction of the least-squares fitting of the theoretical envelope to acquired isotopic clusters and calculation of its 'prior ion' ratio. In this example, the theoretical isotopic envelope of the SILAC-heavy form of the peptide GPCMSEQAMGPCMSEQAMK is fitted to the acquired intensities. A low intensity, 'prior ion' is observable prior to the intense monoisotopic peak. The ratio of the sum of the intensities fitted to the theoretical envelope by least squares fitting (blue bars) divided by prior ion intensity is termed the 'prior ion ratio'. In the first case this ratio is small because no overlapping isotope cluster is present (black sticks), but in the second, an indistinguishable interfering or co-eluting cluster (red sticks) is shown which results in a high prior ion ratio despite a good fit from the least squares method.

possible to calculate the median protein fold changes for a range of least squares cut offs starting at 1.0 and descending to 0 in 0.01 intervals. This was a straightforward task because the protein data was also available directly in the .hdf5 file and it was simple to link the peptides to their corresponding proteins. The median fold changes of the different replicates (on peptide and protein level) were then compared and the IQR was plotted for the difference between the two replicates. In cases of repeat peptides the best scoring PCM was used from both replicates. For this assessment all data passing the described filters were included (i.e. no prior ion filter was applied). The number of quantified PCMs (peptides) passing the cut offs and the least-squares-fitted intensity was recorded as well as the total number of quantified proteins yielded.

2.5.2 Assessment of prior ion ratio cut off

The results of the first 7 hour monocyte data set were treated as described for the least squares threshold (see 2.5.1) and the range of thresholds tested started at 0.5 and descended to 0 in intervals of 0.01. Density plots were made of spread of the $\log_2$ deviation of each individual peptide fold change from the $\log_2$ fold change its protein at the different cut offs.

2.5.3 Assessment of the exclusion of missing peptide fold changes in the calculation of the protein fold change

Following processing as described above and after filtering away peptides based on the resulting optimized thresholds, the effect on protein fold change reproducibility between the 7 hour monocyte replicates was assessed after removing all peptides with indeterminable ratios (fold changes of zero resulting from one missing channel but whose peptide otherwise passes all quantification-dependent filters). The trigger for this removal was different total counts of positive peptide fold changes for each protein starting at a minimum of one and going up to maximum tested value of ten.

2.6 Comparison between isobarQuant and MaxQuant

In order to compare quantification precision and accuracy to a widely used software, a THP-1 SILAC mixture was analyzed with both isobarQuant and MaxQuant[138]. A dilution series was set up and acquired data were processed with MaxQuant (Version1.5.8.0) using the settings generally recommended for SILAC quantification[226]. Searches were performed using the same search database (see paragraph 2.2.0.1) for both softwares using carbamidomethylation of cysteine as a fixed modification and oxidation of methionine and acetyl (protein N terminus) as variable modifications. The mass tolerance for the precursor was 4.5 ppm and 20 ppm for the fragment ions, ‘re-quantify’ option was switched on, (a separate analysis was also performed with the ‘re-quantify’ option switched off), in the section ‘group-specific parameters’ a multiplicity of 2 was selected, with Arg10 and Lys8 chosen as the ‘Heavy labels’. The settings used for quantification were the same as above (2.2.1), but implemented the optimal results from the least squares, prior ion ratio and non-zero filtering investigations.

2.7 Protein half-life determination

Protein half-lives were determined according to a modified version of the protein decay rate method described by Schwanhäusser *et al.*[51]. Because these are non-dividing cell types, the cell cycle time correction component was removed from the equation to give the formula:

$$k_{dp} = \frac{\sum_{i=1}^{m} log_e(r_{t_i} + 1) \cdot t_i}{\sum_{i=1}^{m} t_i^2}$$

where k_{dp} is the rate constant of the protein decay, m is the number of time points (t_i) considered and $r_{t_i}$is the fold change ratio (heavy / light) of a specific protein at each time point. The half-life of a protein ($T_{1/2}$) is then calculated by

$$T_{1/2} = \frac{log_e 2}{k_{dp}}$$

For each protein, a linear model was fitted to the time course of the logarithmic protein fold changes $log_e(r_{t_i} + 1)$ and the coefficient of determination (R^2) for the linear regression was recorded. The QC value was set to 'weak' if it was possible to determine a fold change in at least three out of the four time points, 'good' if the protein fold changes at three out of the four time points were based on a minimum of three quantified peptides and 'poor' for the remainder.

2.8 Assessments of protein half-lives in complexes

2.8.1 Analysis of half-life variability within protein complexes

All proteins identified in any of the cell types were mapped to complexes using the hu.MAP complex database[227]. A Python script was used to filter the complexes to contain at least five quantified protein members in a given cell type and the standard deviation of the $\log_{10}$ transformed half-lives among the complex members was computed. In order to create a reference that indicated half-life variability as expected by chance, random representatives were drawn from the list of proteins associated with the complexes in a given cell type and placed into groups representing the different numbers of true complex members. Thus, in the end, a random group of proteins associated with a complex will have the same number of proteins as the original, true complex group. Once again the standard deviation of the $\log_{10}$ transformed half-lives among the proteins within each group was computed. Differences in the $\log_{10}$ half-lives of true protein complex members vs. the random draws of proteins in a given cell type were assessed by Wilcoxon-rank test (significance levels were encoded as *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$). The results were plotting using Python and its visualization library, Matplotlib.

2.8.2 Mapping of protein half-lives onto protein complex structures

For each cell type, the mean protein half-lives calculated from two replicates were mapped onto protein complex structures as a linear three-color gradient. To avoid distorting the gradient by outliers with exceptionally high or low half-lives, the gradient midpoint was set to the median half-life of the subunits of the complex and the lower and upper half-life values for the linear interpolation were set to the 15th and 85th percentile, respectively. The half-lives outside this percentile range were clipped. The median and percentiles were calculated as a mean of, respectively, the medians and percentiles of each biological replicate. For coloring the Nup214 complex, which is represented in the structure as a single density segment composed of three subunits, the mean half-life of all three subunits was used. All calculations were performed using half-lives with an R^2 of at least 0.25.

3 Results

isobarQuant was adapted to perform MS1-based quantification as part of the first stage of the post-Mascot workflow. Since the SILAC quantification values are linked to the results of the Mascot search (basis for the exact model), the quantification values are stored in the .hdf5 file within the same group as the Mascot results. Together with the determined least squares intensities, the prior ion ratios and least squares fit values are recorded in the .hdf5 file. These are later used to filter the values taken forward for the calculation of the protein group's median fold change. The adaptations also enable isobarQuant to handle other MS1 (precursor) intensity-based quantification methods including dimethyl labeling[58], mTRAQ[228] and hyperplexed dimethyl modes[59,229].

3.1 Averagine versus exact ion model

Compared with averagine, the fits are on average better for the exact model. For the light peptides the mean improvement in -$\log_{10}$ transformed least squares fits is 0.17 and 0.38 for the heavy peptides. The overall stronger improvement for the heavy peptides is brought about by the inclusion of masses of the modified lysine and arginine residues when calculating the theoretical isotopic envelope, while the averagine model does not distinguish between heavy and light peptides and therefore performs worse for heavy peptides. However, the improvements are not universal. In some cases, the least squares fits are poor for both models and no improvement is observed indicating that the measured isotopic clusters do not fit well to either model. For more than one third (32, 494) of all isotopic distributions determined for identified peptides (100, 938) a worse least squares fit was obtained with the exact model compared to the averagine model and these generally correspond to identifications with lower Mascot scores (mean ion score 39 vs. 46).

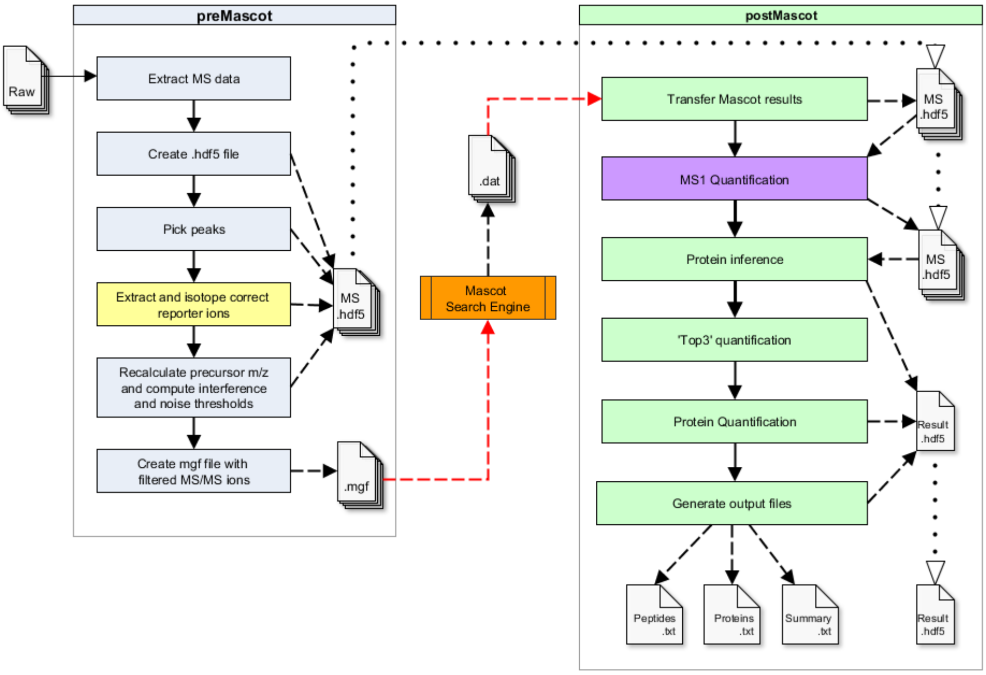


Figure (4): The adapted isobarQuant workflow showing the location of the MS1 quantification step in purple, taken from ref.[181]. The monoisotopic mass of the peptide (plus charge and modification(s)) is used to determine exact mass and isotope distribution from its atomic composition which can then be matched to the observed precursor data from the .raw file also stored within the .hdf5 file.

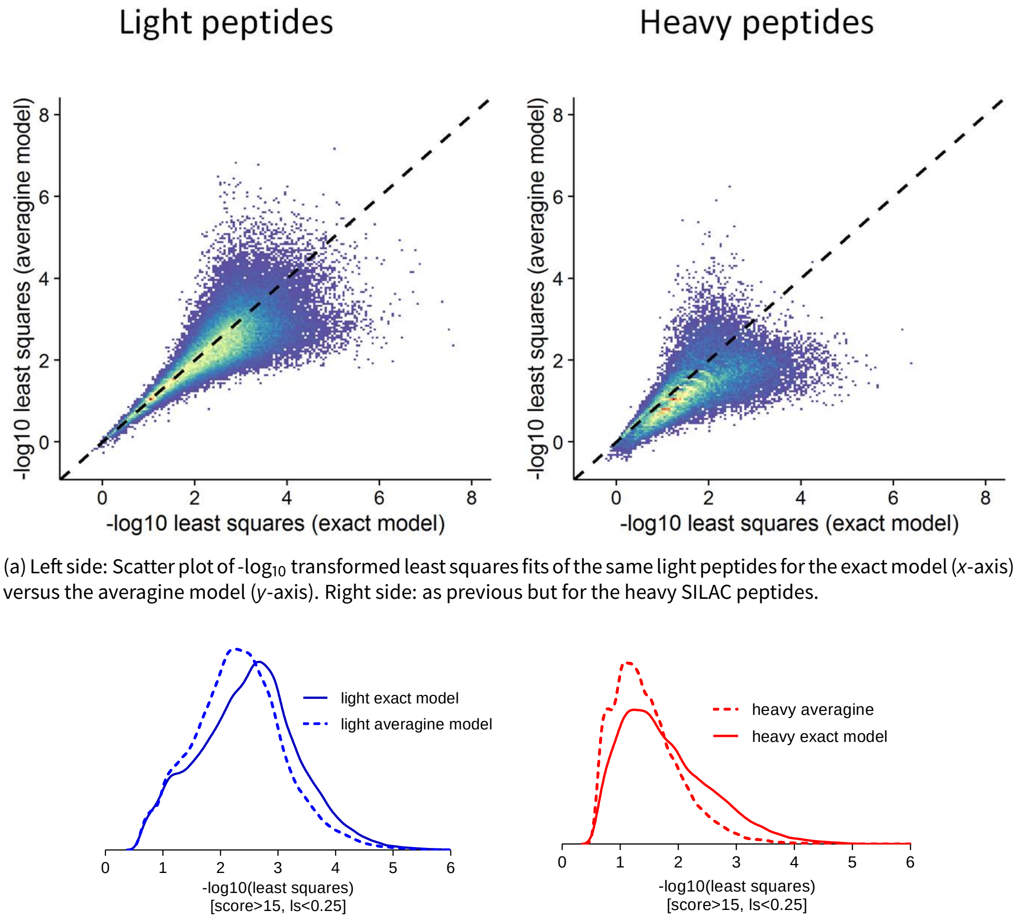
3.2 Optimization of filters for maximum coverage with highest precision and accuracy

The aim of this part of the investigation was to enable more peptides and proteins to be quantified more accurately and with greater precision. The investigation into the effect of using the averagine model was also extended. To this end, the adapted isobarQuant was run in two modes (once using the exact model and once with averagine).

3.2.1 Assessment of peptide filters based on fit quality of isotopic distributions and prior ion ratio for improving quantification accuracy.

A substantial improvement in reproducibility (measured by the interquartile range of the delta $\log_2$ fold change between replicates) compared to all proteins is achieved using the exact model when applying the filter at 0.1, with relatively low loss of proteins. A similar trend is observed for the averagine model up to a least squares cut off of 0.1, however the protein fold changes are less reproducible throughout; after this cut off the reproducibility for the averagine model becomes more unstable.

The exact model starts with around 300 (approximately 5%) fewer quantified peptides than the averagine model; these are all filtered away around the least-squares threshold of 0.5. After this point the exact model consistently yields more peptides than with averagine. The number of peptides retained at the different least squares thresholds above approximately 0.1 decreases slowly, Fig. 6a. After this point the decrease becomes more rapid with the increasingly stringent filtering. Only 5% of the



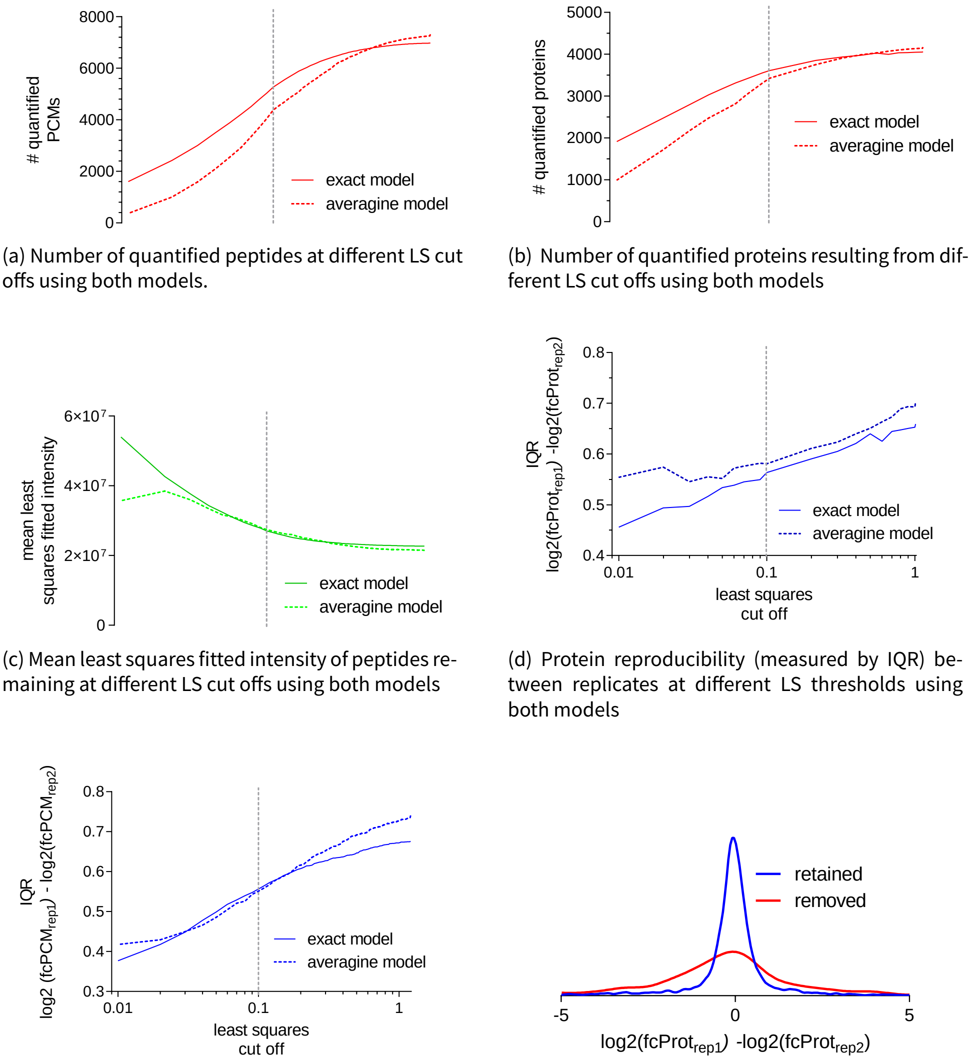
(a) Left side: Scatter plot of -$\log_{10}$ transformed least squares fits of the same light peptides for the exact model (*x*-axis) versus the averagine model (*y*-axis). Right side: as previous but for the heavy SILAC peptides.

(b) Density distribution of the $-\log_{10}$ transformed least squares fits where fit is less than 0.25 and Mascot score for underlying peptide is greater than 15. The dashed line represents the averagine model and the solid line the exact model. Left, light peptides; right, heavy. The improved performance for the exact model indicated by the shift to the right of the density is more marked for the heavy peptides

Figure (5): Data from thirteen .raw files from one cell type (monocytes), harvested at a single time point (7 hours after swapping to heavy medium) was analyzed with isobarQuant, once using an exact model and once with an averagine model.

peptide spectra matched to the averagine model at the most stringent filter criterion (0.01 least-squares fit cut off) compared with 23% achieved using the exact model. The mean signal abundance of matched peptides increases as more stringent cut offs are applied, with the trend more pronounced for the exact model, Fig. 6c. This may be accounted for by the fact that peptides with higher signal intensity have better ion statistics and therefore measured isotope patterns match better with their corresponding theoretical isotopic envelopes. The trend is more pronounced for the exact model than the averagine model and least squares fits for high abundance peptides are markedly better than those achieved with the averagine model (Figs. 6a & 5a). The reproducibility in peptide quantification, measured as the IQR of fold change differences between identical peptides in the biological replicates, is better with more stringent filters. A substantial improvement in reproducibility compared to all quantified peptides is achieved using the exact model; with a least squares filter of 0.1 which retains 75% of all quantified data (Fig. 6e). The averagine model yields the same reproducibility as the exact model at the 0.1 cut off, but substantially fewer peptides are quantified (40% are lost as compared to 25% with the exact model). Similar trends were observed on the protein level (Figs. 6b, 6d & 6f) . At the least squares cut off at 0.1, where 87% of all proteins are still quantifiable but after this point the rate at which proteins are filtered with each least squares cut off increases. At a cut off of 0.1, the reproducibility in protein quantification, IQR, using the exact model is improved by 15% compared to the averagine model, Fig. 6d. The reproducibility of the proteins that are kept after applying the 0.1 threshold using the exact model, and assessed by calculating the standard deviation of the distribution of $\log_2$ protein fold change differences between two biological replicates, is much better, (σ=1.21, Fig. 6f, blue line) than the reproducibility of the proteins that are filtered away, (σ=2.15), Fig. 6f, red line).

The influence of the prior ion ratio, the ratio between the ion intensity of a peak corresponding to the loss of one neutron from the monoisotopic ion and the ion intensity from the peptide of interest, on precision of quantification was investigated for the exact model data. The number of quantified peptides decreases with increasing prior ion ratio, 7a. The fitted intensity increases slightly with more stringent prior ion cut offs up to a value of 0.08, after which it decreases rapidly as the most intense ions are filtered out. This figure coincides with the maximum possible miss-incorporation rate for atoms into the heavy label. At this threshold there is essentially no difference between a potentially interfering cluster and the intensity expected due to a prior ion, so a large number of peptides get filtered out (Fig. 7b). Reproducibility was gauged by measuring the IQR between identical peptides in each of the biological replicates. The improvement in reproducibility is steady until a cut off of 0.08 after which it rapidly worsens as the more intense data, which is typically more robust, is filtered away, (Fig. 7c). The optimal cut off for prior ion ratio appears to lie between 0.08 and 0.2. Since there is a difference of 17% in the number of quantified peptides between these thresholds and no substantial difference in reproducibility, the value



(a) Number of quantified peptides at different LS cut offs using both models.

(b) Number of quantified proteins resulting from different LS cut offs using both models

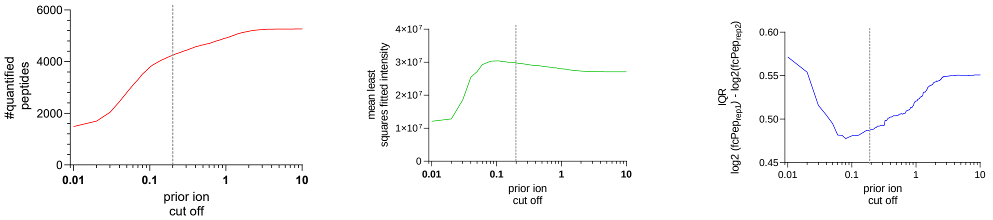
(c) Mean least squares fitted intensity of peptides remaining at different LS cut offs using both models

(d) Protein reproducibility (measured by IQR) between replicates at different LS thresholds using both models

(e) PCM reproducibility (measured by IQR) between replicates at different LS thresholds using both models

(f) Density distributions of $\log_2$ protein fold change difference between biological replicates for proteins removed or retained at 0.1 threshold

Figure (6): Assessing the impact of the least squares fits of labeled peptides The gray dashed line shows the selected threshold: 0.1 for the quality of the least squares fits of peptide isotopic distributions calculated as the sum of the squares of the residual values. Figure (a): Despite starting with 300 more quantified PCMs, the exact model consistently yields more peptides than averagine with least-squares thresholds more stringent than 0.5. The number of peptides retained for the both models decreases slowly above the cut off of 0.1. Only 5% of the starting number of PCMs are retained at the most stringent cut off for the averagine model compared to 23% for the exact model. Figure (b): The trends observed for PCMs (a) are reflected in the numbers of quantified proteins when differnt least squares cut offs are applied, with nearly double the number of quantified proteins being observed for the exact model compared to the averagine model at the most stringent cut off. Figure (c) The mean signal abundance of matched peptides increases as more stringent cut offs are applied, with the trend more pronounced for the exact model, which is likely because higher signal intensities have better ion statistics and consequently better matches to the theoretical isotopic envelope. Figure (d): Protein reproducibility is consistenly better for the exact model at all least-squares cut offs with a 15% improvement over the averagine model at the 0.1 threshold. Figure (e): PCM reproducibility is either similar between the two models or better with the exact model at all least squares cut offs. Figure (f): For the exact model, the standard deviation of the distribution of reproducibility between the two biological replicates (defined as the difference in $\log_2$ protein fold change) of proteins that are kept after applying the 0.1 threshold is 1.21; much better compared to 2.15 for those proteins excluded at this threshold.



(a) count of quantified PCMs at different prior ion cut offs

(b) mean intensity at different prior ion cut offs

(c) IQR of peptides at different prior ion cut offs

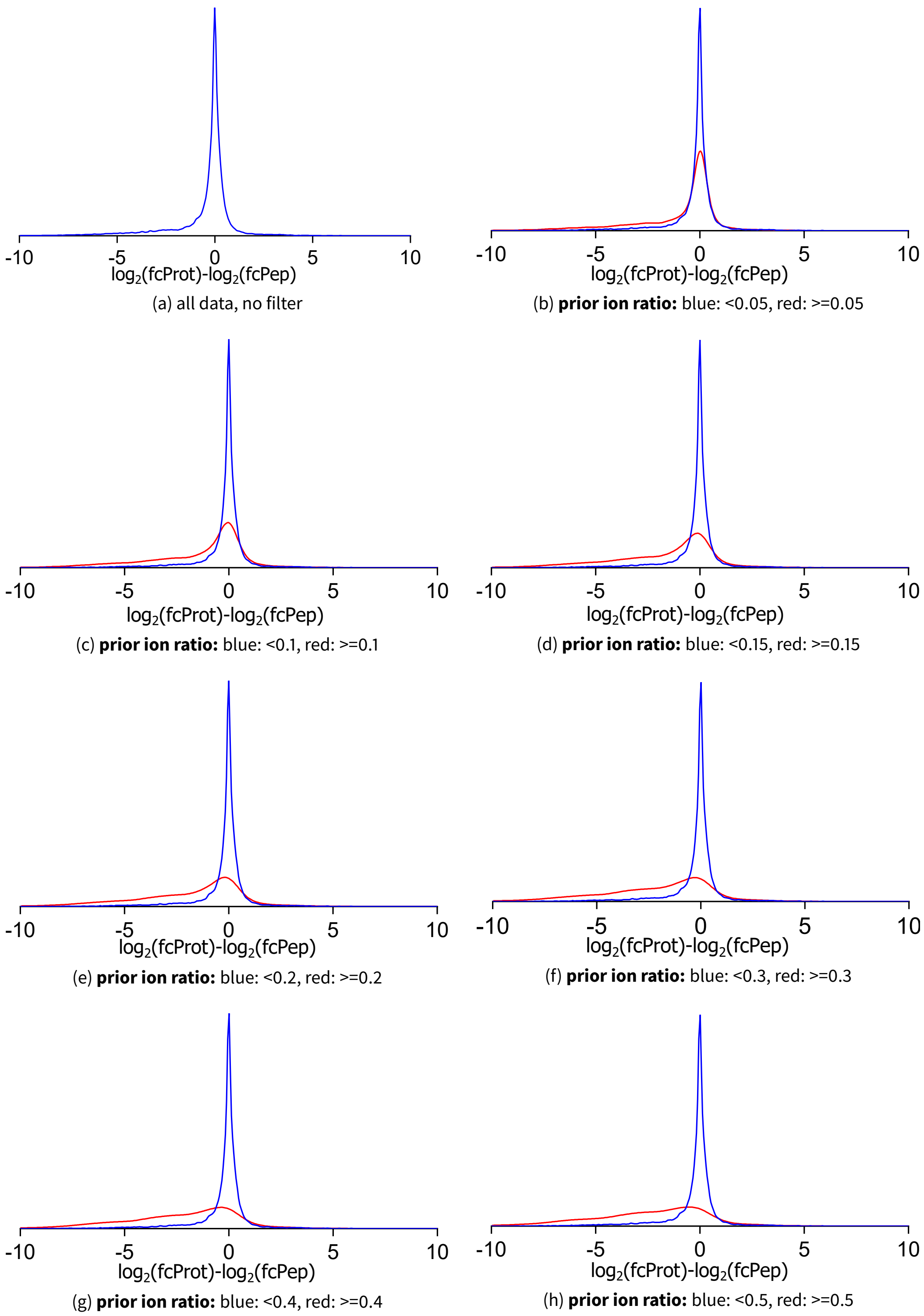
Figure (7): The effect of different prior ion thresholds on accuracy and precision of quantification. Figure (a): The number of quantified peptides decreases uniformly with increasingly stringent cut offs dropping slightly as the cut off approaches 0.01. Figure (b): The calculated fitted intensity of quantified peptides increases slightly with more stringent prior ion cut offs up to a value of 0.08, after which it decreases rapidly as the most intense ions are filtered away. This figure coincides with the maximum possible miss-incorporation rate for atoms into the heavy label. At this threshold there is essentially no difference between a potentially interfering cluster and the intensity expected due to a prior ion. Figure (c): Reproducibility measured as the IQR of the delta of quantification values of identical peptides in the two replicates.The IQR improves with increasingly stringent prior ion criteria until the threshold of 0.08, after which it rapidly worsens. This is again due to the higher-intensity peptides (and their robust quantification values) being filtered away.

of 0.2 was chosen for the prior ion ratio cut off (show as gray dashed line in Fig. 7). This retained approximately 80% of all quantified peptides and provided a substantial improvement in reproducibility. The effect of prior ion filtering on reproducibility is much less marked than that of least squares filtering, which can be explained by the fact that two biologically-similar (here technical replicates) samples are likely to produce the same peptides eluting at a very similar time. Despite the data being relatively reproducible between replicates it is still possible that the accuracy is impaired by other interfering peptides.

As depicted in figure 3, the presence of an intense 'prior ion' should not only affect precision but also the accuracy of the heavy to light ratio. The presence of interfering isotopic clusters can be partially mitigated by least squares filtering of peptides as described above, but acceptable matches may still contain interfering signals and inaccurate ratios. To investigate this, we estimated the accuracy of individual peptide ratios by measuring the difference between their $\log_2$ fold change and the $\log_2$ median fold change of the corresponding protein group when different prior ion ratio thresholds were applied. Figure 8 shows that for peptides with a prior ion ratio of greater than 0.2 there is a greater spread, corresponding to lower accuracy, in the data and is summarized in figure 9. The asymmetric nature of the plots (the broader left shoulder) arises because of the underestimation of peptide fold changes which, in this data set, occurs more often because the signal of the heavy channel is more likely to be increased due to the presence of interfering ions, while the interference in the light channel is much less likely. This is because at the early time point of seven hours, the light signal is more abundant than the heavy signal and is thus, in most cases, less susceptible to influence from prior ions that would lead to an overestimation of the peptide's heavy to light ratio. This phenomenon is analogous to the ratio compression observed with TMT and iTRAQ isobaric labeling[169,171,230], where the reporter ion fold changes are altered by contaminating reporter ion signals from co-eluting peptides. Figures 10a and 10b showcase a couple of example proteins where the prior ion ratio is high and the peptide fold change is inaccurate. This fact, combined with the modest gains in precision, means that using a filtering value of 0.2 for the prior ion ratio is a good compromise between a modest loss in quantified peptides and the improved accuracy and precision.

3.2.2 Impact of missing values on protein fold change determination

Protein fold changes based on precursor intensities are usually derived from the median value of the ratios of all peptides linked to that protein passing certain QC criteria. The median should minimize the effect any extreme peptide fold changes have on the result. However, there are cases where it is not possible to determine a peptide fold change (for instance where the signal was not detected for one of the labels) but where the peptide otherwise passes all QC criteria. When there are many such cases it can incorrectly lead to a protein fold change of zero or infinity. Imputation, where missing values are replaced with zeros (under the assumption that they are missing because of low abundance), has been shown to perform sub-optimally and can lead to a bias in label-free quantification experiments for moderately and highly abundant proteins, an observation that also holds for SILAC based MS1 quantification[219]. To circumvent this, isobarQuant was adapted to be able to remove the peptides with indeterminable ratios provided that at least a given number of finite ratios was present. Figure 11a shows the effect of removing the peptides with indeterminable ratios on the fold change reproducibility for proteins binned according to number of quantified peptides. Reproducibility was once again inferred using the IQR between identical protein groups in each of the two biological replicates in each bin. Removing indeterminable ratios leads to greater reproducibility as indicated by smaller IQRs. The fold change reproducibility is also more stable across all bins. Using one measured peptide ratio as the trigger to remove indeterminable peptide ratios gives the most reproducible results as seen by profiling different trigger values (Fig. 11). The beneficial effect of removing indeterminable peptide ratios was observed to positively affect the numbers of proteins for which a half-life was determined, not just in the test set, but for all cell types (Fig. 11b). In figures 11c & 11d we see specific examples of the outcome of removing these indeterminable peptide ratios where the remainder yield protein fold changes of 0.017 and 0.069 respectively, indicated by the slope of the red line. In cases where the peptides with indeterminable ratios are left in, a protein fold change cannot be established using the median. This setting of minimally one finite fold change before removing all infinite values was chosen as a default for peptide ion intensity based quantification in isobarQuant.



(a) all data, no filter

(b) **prior ion ratio:** blue: <0.05, red: >=0.05

(c) **prior ion ratio:** blue: <0.1, red: >=0.1

(d) **prior ion ratio:** blue: <0.15, red: >=0.15

(e) **prior ion ratio:** blue: <0.2, red: >=0.2

(f) **prior ion ratio:** blue: <0.3, red: >=0.3

(g) **prior ion ratio:** blue: <0.4, red: >=0.4

(h) **prior ion ratio:** blue: <0.5, red: >=0.5

Figure (8): Effect of prior ion ratio on fold change accuracy: Density plots of the spread of peptide fold-change deviations (determined by subtracting the $\log_2$ fold change ratio of the peptide from the $\log_2$ fold change [median] of the corresponding protein; *x*-axis: log2(fcProt)-log2(fcPep)) for data retained or filtered out at increasingly stringent cut offs. The plot is skewed for data filtered away at cut offs greater than 0.15, indicating that there is a consistent overestimation of peptide fold changes in these data. At a prior ion ratio cut off of 0.15 and below, the spread of the data is narrow, at higher cut offs the spread is substantially wider.

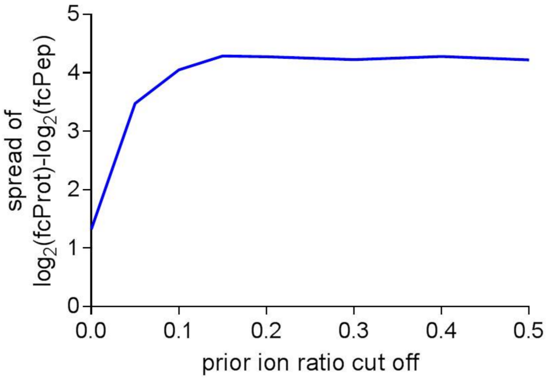
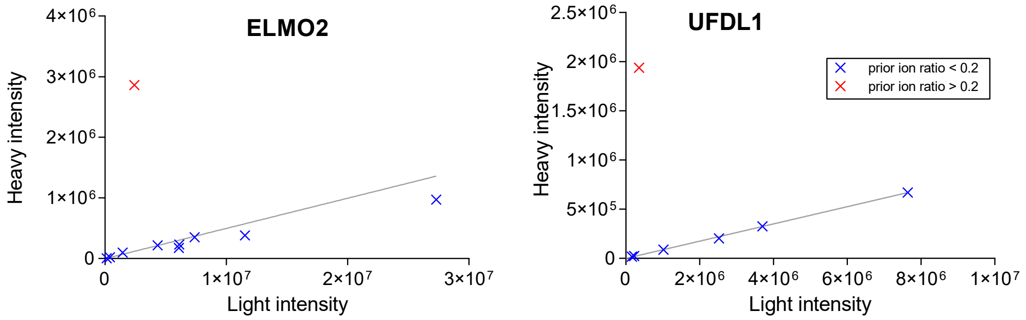


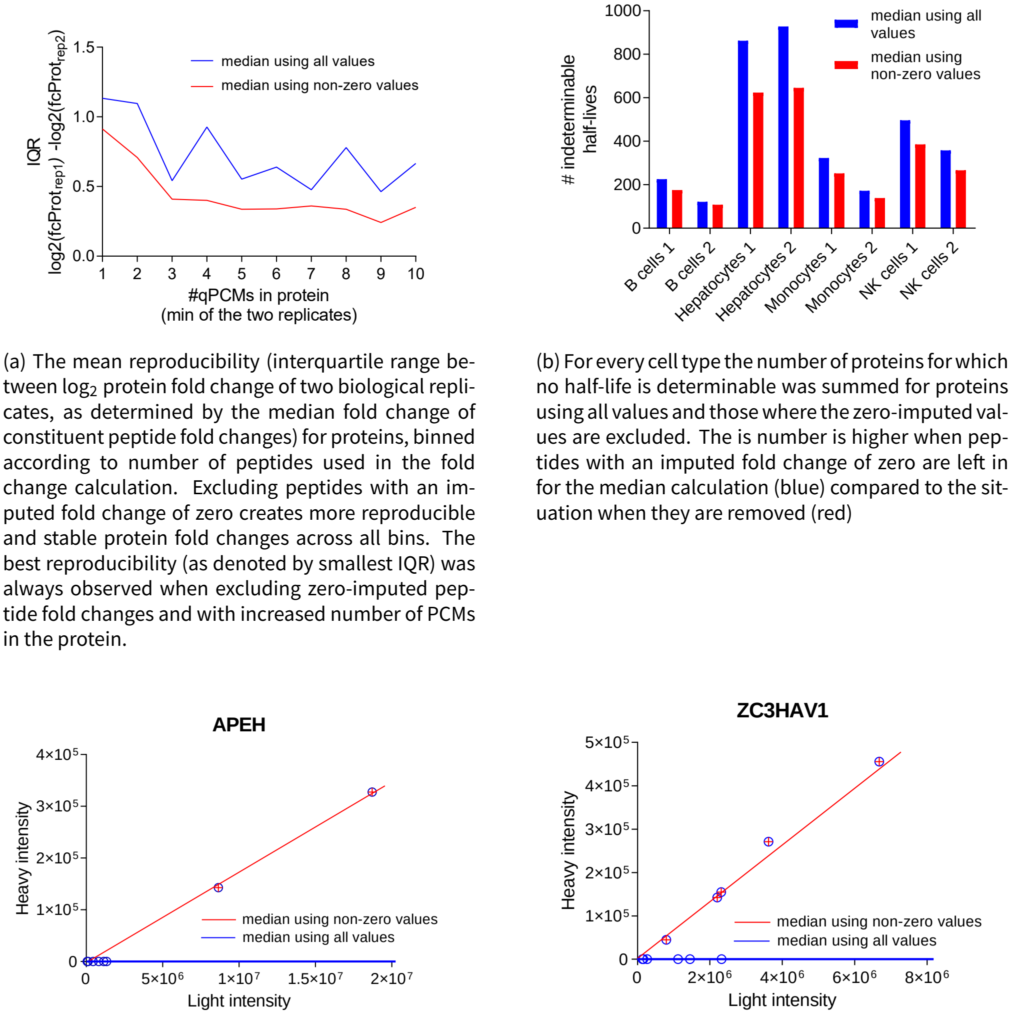
Figure (9): Summary of the effect of prior ion ratio on fold change accuracies confirms that the spread is greater in data with a prior ion ratio above 0.15. The ratio for each peptide was calculated by subtracting the log transformed peptide fold change from the median log transformed fold change of the corresonding protein. The mean spread of all deltas is plotted against the different prior ion ratio cut offs.



(a) For ELMO2 the single outlying peptide (red cross) has a fold change of 1.2 (compared to the median fold change of 0.05) and a prior ion ratio of 3.002, with all other peptides having a prior ion below 0.2 (blue crosses)

(b) For UFDL1 the outlying peptide (red cross) has a fold change of 5.4 compared to the median fold change of 0.09 for the protein. It has a prior ion ratio of 1.194 which contrasts to all other peptides (blue crosses) which have a prior ion ratio below 0.2.

Figure (10): Effect of prior ion ratio on of fold change accuracies of peptides for two selected, individual proteins. A prior ion ratio greater than 0.2 is a key indicator that the fold change of the given peptide is not in line with that of other peptides in the same protein.



(a) The mean reproducibility (interquartile range between $\log_2$ protein fold change of two biological replicates, as determined by the median fold change of constituent peptide fold changes) for proteins, binned according to number of peptides used in the fold change calculation. Excluding peptides with an imputed fold change of zero creates more reproducible and stable protein fold changes across all bins. The best reproducibility (as denoted by smallest IQR) was always observed when excluding zero-imputed peptide fold changes and with increased number of PCMs in the protein.

(b) For every cell type the number of proteins for which no half-life is determinable was summed for proteins using all values and those where the zero-imputed values are excluded. The is number is higher when peptides with an imputed fold change of zero are left in for the median calculation (blue) compared to the situation when they are removed (red)

(c) APEH protein for which the inclusion of peptides with a measured fold change of zero results in an indeterminable protein fold change (blue) compared to the case when they are not used (red). The median is represented by the slope of the line.

(d) ZC3HAV1 protein for which the inclusion of peptides with a measured fold change of zero results in an indeterminable protein fold change (blue) compared to the case when they are not used (red). The median is represented by the slope of the line.

Figure (11): The effect of excluding peptides with a ratio of imputed-zero on the protein fold change calculation when there is at least one measured peptide ratio

mixing ratio	$\log_2$ fold change equivalent
1:1	0 $\log_2$ ratio
1:9	3.2 $\log_2$ ratio
1:49	5.6 $\log_2$ ratio

Table (6): Table giving mixing ratios for (heavy to light) SILAC-labeled cells for testing accurcy of isobarQuant and MaxQuant.

3.3 Comparison to existing software (MaxQuant)

With the newly-adapted isobarQuant software in place and the parameters optimized, the next step was to test it against another software which performs peptide ion based quantification – MaxQuant[138], which uses the averagine model and inbuilt filter criteria to exclude peptides from the protein fold change calculation.

Light and heavy SILAC labeled THP1 cells were mixed at different ratios, and the lysed, digested sample was measured without any pre-fractionation thus creating a particularly demanding task for accurate quantification. The ratios are given in table 6, with the most demanding ratio being a 1:49 mix of light to heavy SILAC cells

The deviation in accuracy of the mode (the most frequently occurring value in the distribution) for the 1:1 mix is representative for the deviation due to pipetting precision. I compared the performance of isobarQuant to MaxQuant using both the widely used and recommended setting of re-quantify “on” as well as the re-quantify setting turned off. Re-quantify attempts to rescue and find new peptide signals by looking in the relevant retention time window for peaks that would fit into the expected isotopic pattern. This function significantly increases the number of accurately quantified peptides, but comes at a cost of including a high amount of less well quantified ones. The isobarQuant strategy for going back and quantifying peptides after peptide identification, is conceptually similar to MaxQuant’s re-quantify function. This strategy evaluates the quality of each theoretical and experimental isotope cluster match and is able to extract then accurately quantify peptides significantly better than MaxQuant when large ratios are measured, which is apparent from the more accurate median value for the 1:49 sample, the addition of filter criteria based on prior ion and isotopic fit quality filters away the poorly quantified peptides and further improves the median value.

In figure 12 row A we first of all see a comparison of the quantification performance on the peptide level when using the MaxQuant ’re-quantify’ function (solid line) and when this function is turned off (dashed line). Without the re-quantify function MaxQuant is able to retrieve accurate quantification values but there is a significant drop in peptides quantified. While this reduces the number of inaccurately quantified peptides, a substantial portion of the accurately quantified peptides is also lost. In figure 12 row B we see that running isobarQuant without any filtering retrieves as many accurately quantified peptides as MaxQuant but significantly fewer poorly-quantified ones. This is immediately clear from the median values of the 1:49 sample (-4 for MaxQuant and – 4.8 for isobarQuant). Figure 12, row C: isobarQuant with the filtering criteria established in this study (mean least squares fit < 0.1 and prior ion

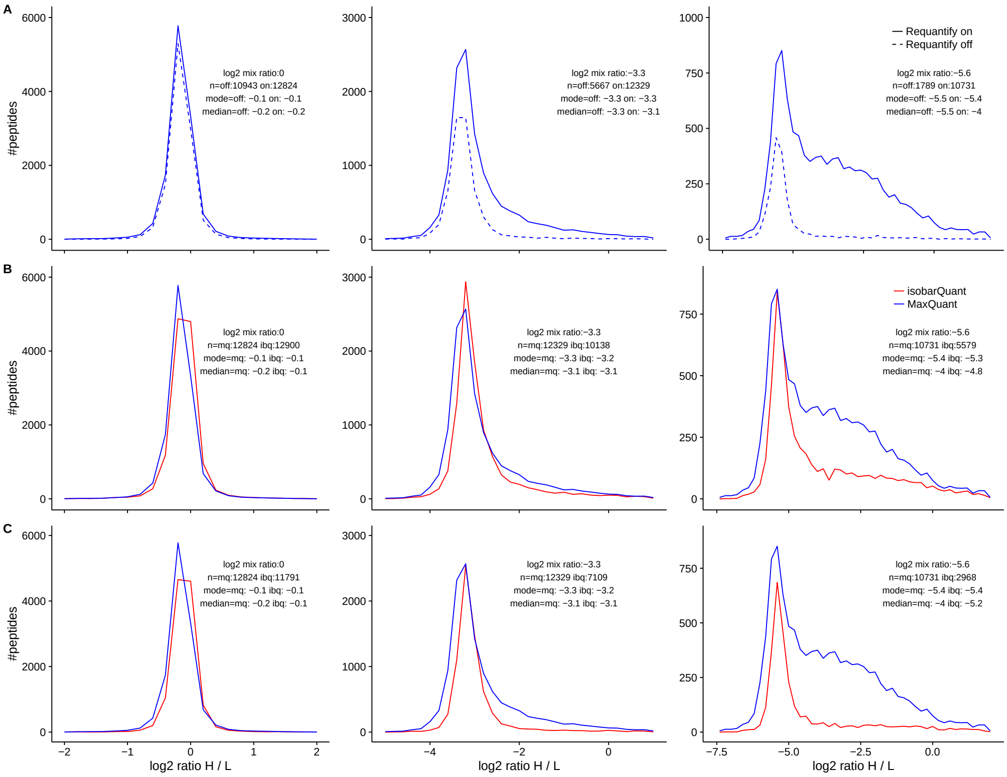


Figure (12): Assessment of peptide ratios obtained using isobarQuant and MaxQuant for three dilutions of heavy SILAC in THP1 cells with decreasing amount of heavy signal from left to right. Column 1: 0 $\log_2$ ratio; Column 2: 3.2 $\log_2$ ratio; Column 3: 5.6 $\log_2$ ratio. Panel (A) compares the performance of peptide level quantification in MaxQuant using the 're-quantify' function (solid blue line) with the default setting (when it is switched off, dashed blue line). Not using the re-quantify function MaxQuant is able to retrieve accurate quantification values but at the cost of a significant drop in quantified peptides. Panel (B): isobarQuant without any filtering steps (solid red line) retrieves as many accurately quantified peptides as MaxQuant but significantly fewer poorly quantified ones. This is immediately clear from the median values of the 1:49 sample (-4 for MaxQuant and – 4.8 for isobarQuant). Panel (C): isobarQuant, with the filtering steps used in this study (means least squares fit < 0.1 and prior ion ratio < 0.2) applied, retrieves a large number of accurately quantified peptides and manages to discriminate and exclude the poorly quantified peptides, which is reflected in a median value of -5.2 in the 1:49 sample.

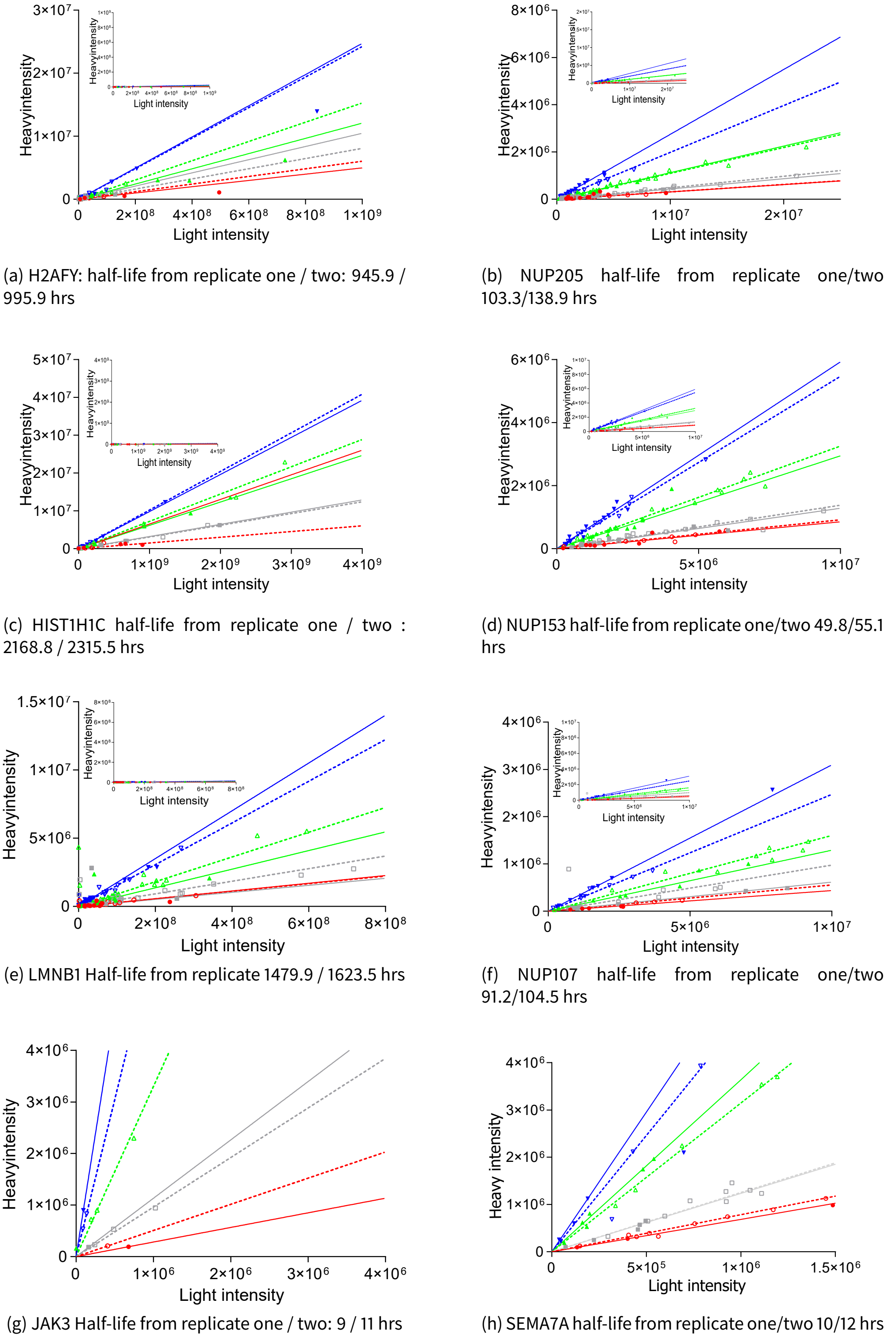
ratio < 0.2) retrieves a large number of accurately quantified peptides and manages to discriminate and exclude the poorly quantified peptides, which is reflected in a median value of -5.2 in the 1:49 sample.

3.4 Protein half-lives in five primary cell types

All biological replicates from all five different non-dividing cell types, human B-cells, monocytes, NK cells, hepatocytes and mouse embryonic neurons (acquired in 569 .raw files) were now processed through the isobarQuant pipeline run using the optimized parameters investigated above. Taken together across all five cell types, protein half-lives were determined for a total of 9,699 unique protein groups. The coverage in individual cell types ranged from 4,667 protein groups identified in NK cells to 6,534 protein groups identified in mouse neurons. To dig deeper into differences between the different cell types a high-quality subset of the data was created by selecting protein half-lives where the regression line (the rate constant of the protein degradation) fitted to fold changes at different time points had a coefficient of determination R^2 of >0.85[51,231]. This criterion was fulfilled by 8,804 proteins across all five cell types. The mean half life was calculated when proteins were present in both replicates for each cell type and these values were used to compare the protein half-lives between cell types, Fig. 14. At the same time, protein half-lives between biological replicates reveal excellent reproducibility with many proteins showing half-lives of greater than 500 hours (Fig. 14). The R^2 of the $\log_{10}$-transformed half-lives between replicates was 0.94, 0.92, 0.91, 0.93, and 0.93 for B-cells, monocytes, NK cells, hepatocytes, and mouse embryonic neurons, respectively. For 98% of all protein half-lives in the high-quality data set, the replicates differed by less than two-fold. The protein turnover in NK cells was the slowest, with 210 proteins having high quality (R^2 of >0.85) half-lives longer than 500 hours. The other cell types (monocytes, hepatocytes, B-cells, and mouse embryonic neurons) had, respectively, only 7, 17, 15, and 4 such proteins. Despite having an overall slower turnover rate, the relative $\log_{10}$ transformed half-lives between NK cells and B-cells, as well as NK cells and monocytes were in good agreement: R^2 of 0.65 and 0.63, respectively. The same holds true for monocytes and B-cells: R^2 of 0.56. In contrast, hepatocytes, which are not of hematopoietic lineage, showed the weakest correlation among the human cells, R^2 of 0.36, 0.41, and 0.36 compared to B-cells, NK cells and monocytes respectively. Half-lives determined in the mouse embryonic neurons agreed slightly better with B-cells, NK cells, and monocytes than with hepatocytes (R^2 of 0.422, 0.567, 0.413 compared to 0.398). Among the fast turnover proteins, we find members of the Janus family of kinases. In particular Janus kinase 3, which is predominantly expressed in the cells of hematopoietic lineage has a very short half-life between 9 and 11 hours (Fig. 13g) in B-cells. Within the longest-lived proteins that were reproducibility observed in more than one cell type, we find the two histone family proteins: HIST1H1C and H2AFY. The average half-life in B-cells is 2,242 and 971 hours, respectively (Fig. 13c & 13a). This value goes up to 2,741 and 1,950 hours, respectively, in NK cells. Interestingly, these histone proteins showed a very quick turnover in hepatocytes; their half-life was 18 and 61 hours, respectively. Lamin-B1 has an average half-life in B-cells of 1,552 hours (Fig. 13e) and in NK cells of 3,215 hours, the half-life in hepatocytes was faster, 388 hours. In mouse embryonic neurons histone HIST1H1B has the slowest turnover, with an average half-life of 1,736 hours, or 72 days. This agrees well with a previous study that used radioactive labeling of the long-lived cerebral histone fractions in mice and reported half-lives of 50–100 days[232].

3.5 Protein half-lives in context – within different complexes

The extensive data set that had been generated using isobarQuant consisted of nearly 10,000 protein half-lives. Since proteins often function within the context of a particular complex, this half life data offered an opportunity to assess the turnover of protein complexes on a proteome-wide scale and to look for coherent turnover behavior



(a) H2AFY: half-life from replicate one / two: 945.9 / 995.9 hrs

(b) NUP205 half-life from replicate one/two 103.3/138.9 hrs

(c) HIST1H1C half-life from replicate one / two : 2168.8 / 2315.5 hrs

(d) NUP153 half-life from replicate one/two 49.8/55.1 hrs

(e) LMNB1 Half-life from replicate 1479.9 / 1623.5 hrs

(f) NUP107 half-life from replicate one/two 91.2/104.5 hrs

(g) JAK3 Half-life from replicate one / two: 9 / 11 hrs

(h) SEMA7A half-life from replicate one/two 10/12 hrs

Figure (13): Plot of peptide fold changes changes for all time points in B-cell identified proteins. The median fold changes are denoted by the slope of the line. The different time points are encoded by the color - red: 7 hours; gray: 11 hours; green: 24 hours, blue: 34 hours, with the individual replicates shown for in either dashed or solid lines. Where displayed, the insets show the SILAC heavy and light intensities plotted on the same scale. H2AFY (13a), HIST1H1C (13c) & LMNB1 (13e) are very long-lived proteins. Two quickly turned-over proteins JAK3 (13g) and SEMA7A (13h) are displayed in the lowest two panels illustrating that a broad range of protein half-lives were recorded for the cell type.

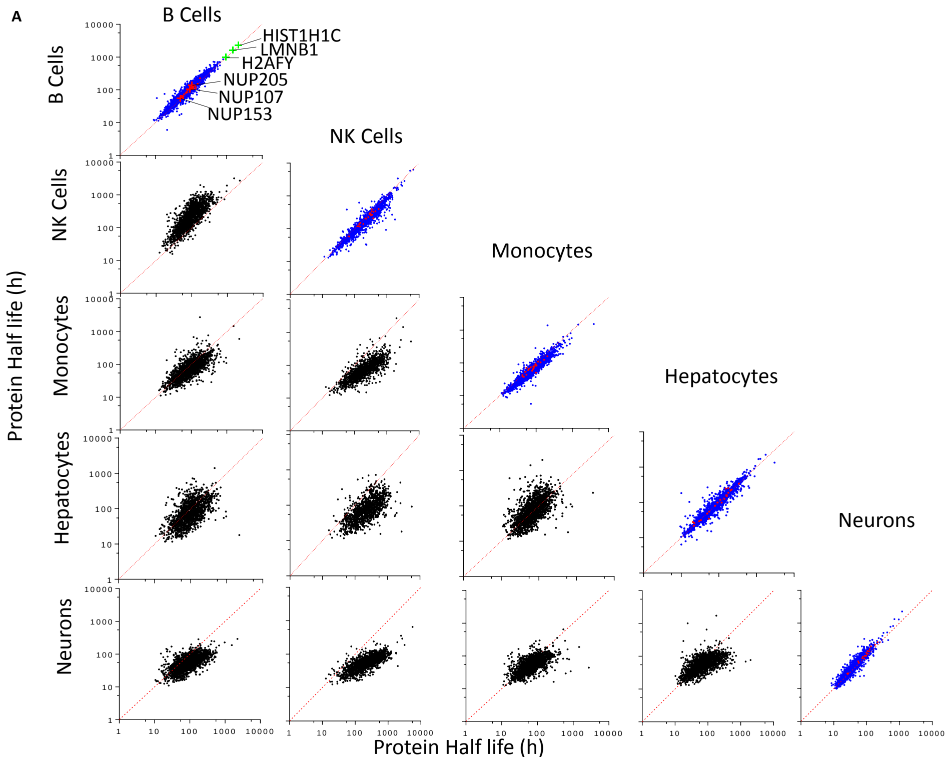


Figure (14): Scatter plot of protein half-lives across and within the different cell types. Plots on the diagonal are the half-lives as determined in each biological replicate for the same cell type. Red dots are members of the nuclear core complex and for the B-cells, green dots mark very long lived histone and lamin proteins, as detailed above (H2AFY: Fig. 13a; LMNB1: Fig. 13e; and HISTH1C: Fig. 13c). The plots below the diagonal compare the mean protein half-lives determined in one cell type against other cell types (as shown).

of proteins located within the same complex. To achieve this, the standard deviation between the $\log_{10}$ half-lives of each complex was calculated and compared to standard deviations calculated for the $\log_{10}$ half-lives of complexes of the same size comprising of randomly allocated proteins. A clear and significant trend ($p\text{-value} < 0.001$, Wilcoxon-rank test) towards a more coherent half-life distribution of protein subunits within individual complexes becomes apparent for all cell types (Fig. 15). The most coherent half-life behavior is observed globally for the nine-member chaperonin complex. Two other large complexes with a more intricate architecture, the nuclear pore complex (NPC) and the 26S proteasome exhibit much less tightly controlled turnover across all cell types, with the exception of the proteasome in mouse neurons, where we see it is among the top 15 complexes with the most coherent behavior.

3.5.1 Proteasome

Looking in greater detail at the 26S proteasome we see significantly different half life behavior between the 20S core complex subunits and the 19S regulatory complex subunits in all cell types (Fig. 17a, 17b), again with the exception of embryonic mouse neurons, which could explain the greater coherence of the protein complex members observed above. Intriguingly, a significant trend for members of the 20S core complex to be more stable than the 19S regulatory complex in B-cells, monocytes and NK cells is observed but a clear and significant trend in the opposite direction is found in hepatocytes (Fig. 17b). Using the root mean square error to estimate the similarity between mean half-lives of all proteasome subunit proteins across all human cell types reveals a distinct separation between the core and regulatory subunits (Fig. 16) and uncovers that PSMD4 and a recently discovered regulatory subunit ADRM1 form a distinct cluster.

3.5.2 Nucleoporins

Another large complex whose half-lives were determined by isobarQuant is the NPC which was the focus of the next part of the investigation. For most cell types, the protein half-lives of these nucleoporins (Nups) reside in the middle of the distribution of all half-lives for all cell types, with a turnover at least one order of magnitude faster than the histone proteins (Fig. 14). The exception is again mouse neurons, but despite showing a wider protein half-life distribution with more, slowly turned over nucleoporins, their overall turnover is still much faster than histones (Fig. 18). In this data set we do not observe differences between the inner ring and Y-complex members; in B-cells the majority of all members turn over just above 100 hours in both subcomplexes. In our comprehensive data set, we do, however, observe a general clustering of half-lives into known subcomplexes (Fig. 19a). The half-lives of members of the Nup358 proteins, and to some extent also the Nup214 proteins, are generally shorter when compared to the inner ring and Y-complexes. The half-lives of members of the Nup62 set, although spatially positioned inside the inner ring complex, appear to be uncoupled from the latter. In hepatocytes and monocytes it is more short-lived but in B-cells more long-lived when compared to other inner ring Nups (Fig. 19a). Interestingly, the turnover of Nup188 is in line with those of the Nup214 complex and Nup98, and an association of which has been proposed. Partitioning the nucleoporins into a scaffold and a peripheral group and comparing the half-life distributions between the two groups shows a statistically significant trend towards faster turnover of the nucleoporins in the peripheral group for all cell types (Fig. 19b). In agreement with previous work[233], we find that Nup98 turns over considerably more quickly than Nup96, although both proteins are synthesized as a single fusion protein prior to autoproteolytic cleavage. This might be explained by the existence of an additional transcript encoding only Nup98. Nup153, Nup50, and the transmembrane Nup gp210 have been shown to have short mean residence times at the NPC[234], although this does not mean that they necessarily turn over once they dissociate. Interestingly, both Nup153 and Nup50 have relatively short half-lives, e.g. 50–70 hours in B-cells. In striking contrast, gp210 generally persists at least as long as scaffold Nups, e.g. ~230 hours in B-cells.

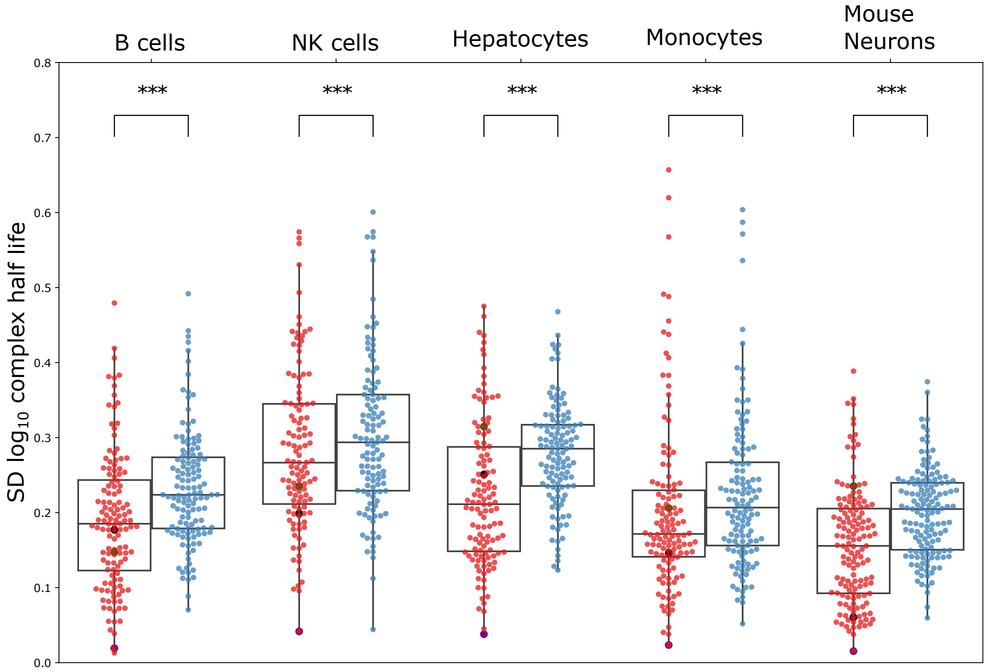


Figure (15): Half-life variability among members of protein complexes is smaller than can be expected by chance. Distributions of standard deviations (SD) of half-lives from proteins in complexes as annotated in the CORUM database (red) compared to standard deviation (SD) of the half-lives of the same proteins shuffled across the different complexes, while preserving the number of proteins in each complex group (blue). Differences in the $\log_{10}$ half-lives of true protein complex members vs. the random draws of proteins in a given cell type were assessed by Wilcoxon-rank test (significance levels were encoded as *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$). Center line in box plots is the median, the bounds of the boxes are the 75 and 25% percentiles i.e., the IQR and the whiskers correspond to the highest or lowest respective value or if the lowest or highest value is an outlier (greater than 1.5 * IQR from the bounds of the boxes) it is exactly 1.5 * IQR

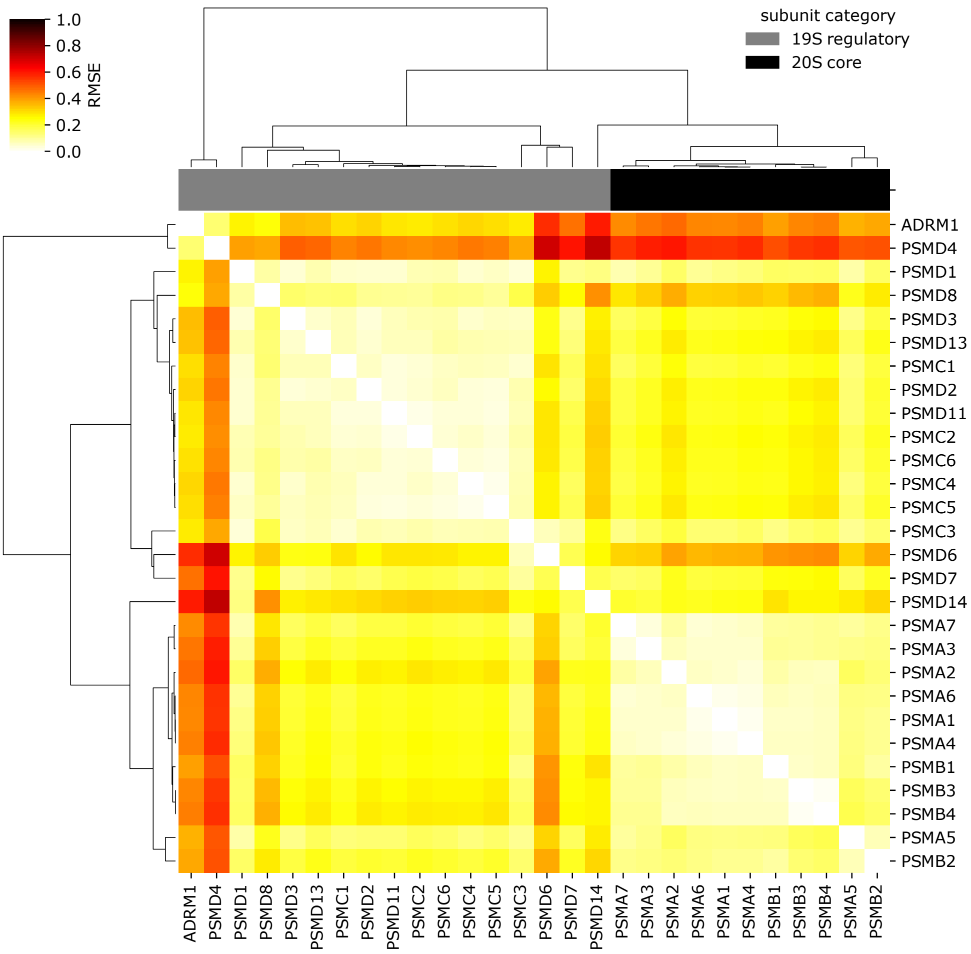
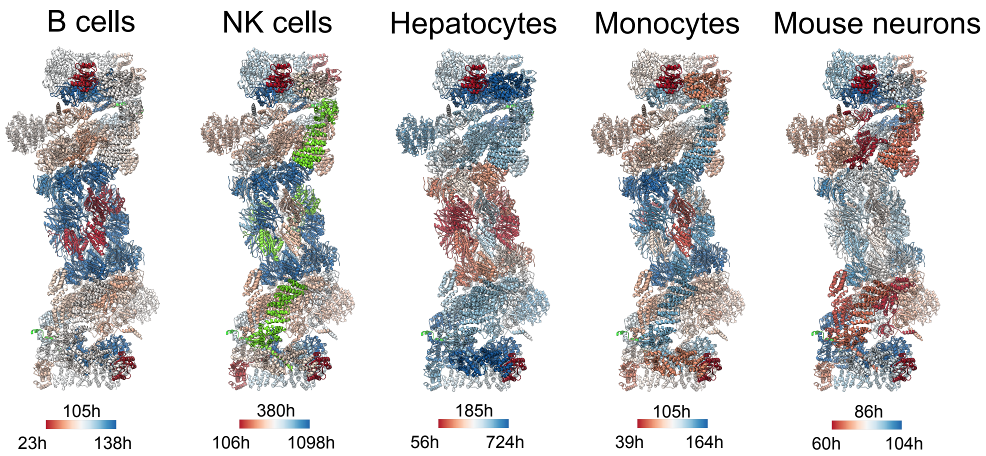
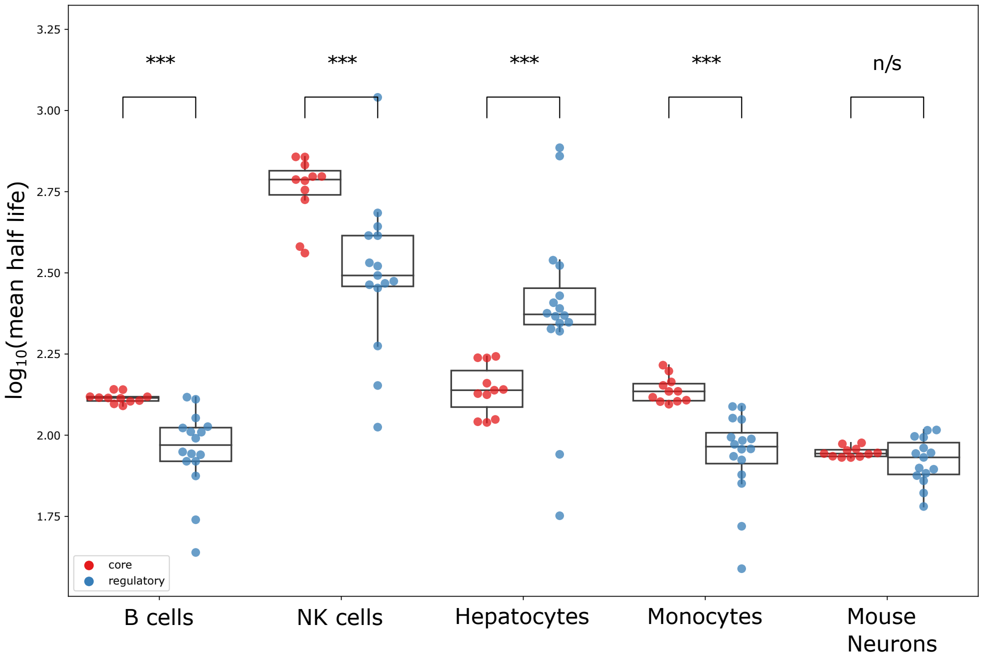


Figure (16): Heatmap showing the comparison for each pair of proteasome subunits by calculating the root mean square error between the four $\log_{10}$ transformed half-lives in the four different human cell types. Hierarchical clustering leads to separation of the regulatory subunits from the non-exchangeable core subunits. The 19S proteasome subunits PSMD4 and the recently discovered ADRM1 also form a distinct cluster



(a) Protein half-lives mapped onto proteasome architecture. Half-lives are depicted by a color gradient ranging from red (shorter half-life) to blue (longer half-life). Median, maximum and minimum half-lives are indicated above, to the right and to the left of the bar. The 20S core subunits stand out as containing more longer-lived proteins than the 19S regulatory complex in all cell types except hepatocytes where this trend is reversed and mouse neurons where there is no significant difference between the two.



(b) Differences in $\log_{10}$ mean protein half-lives for the proteasome in the different cell types. A significant (Wilcoxon rank-sum test - *** $p < 0.001$) difference in mean $\log_{10}$ half-lives between the 20S core subunit proteins (red) and 19S regulatory subunit proteins (blue) is observed in all human cell types. There is no significant difference in the turnover of proteasome proteins in mouse neurons.

Figure (17): Half-lives and architecture of the proteasome

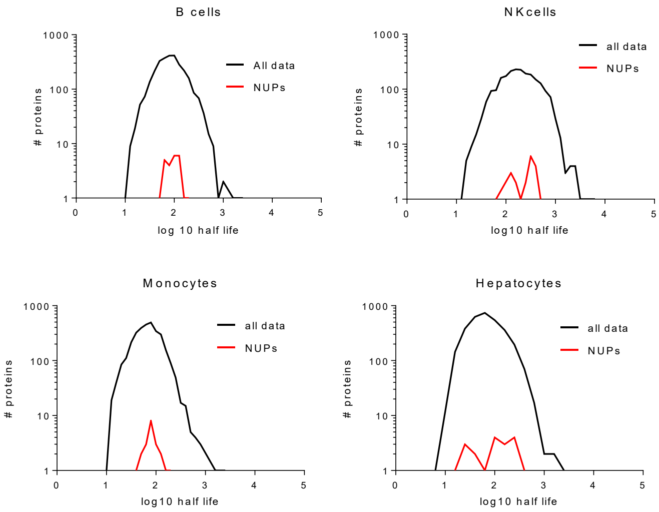
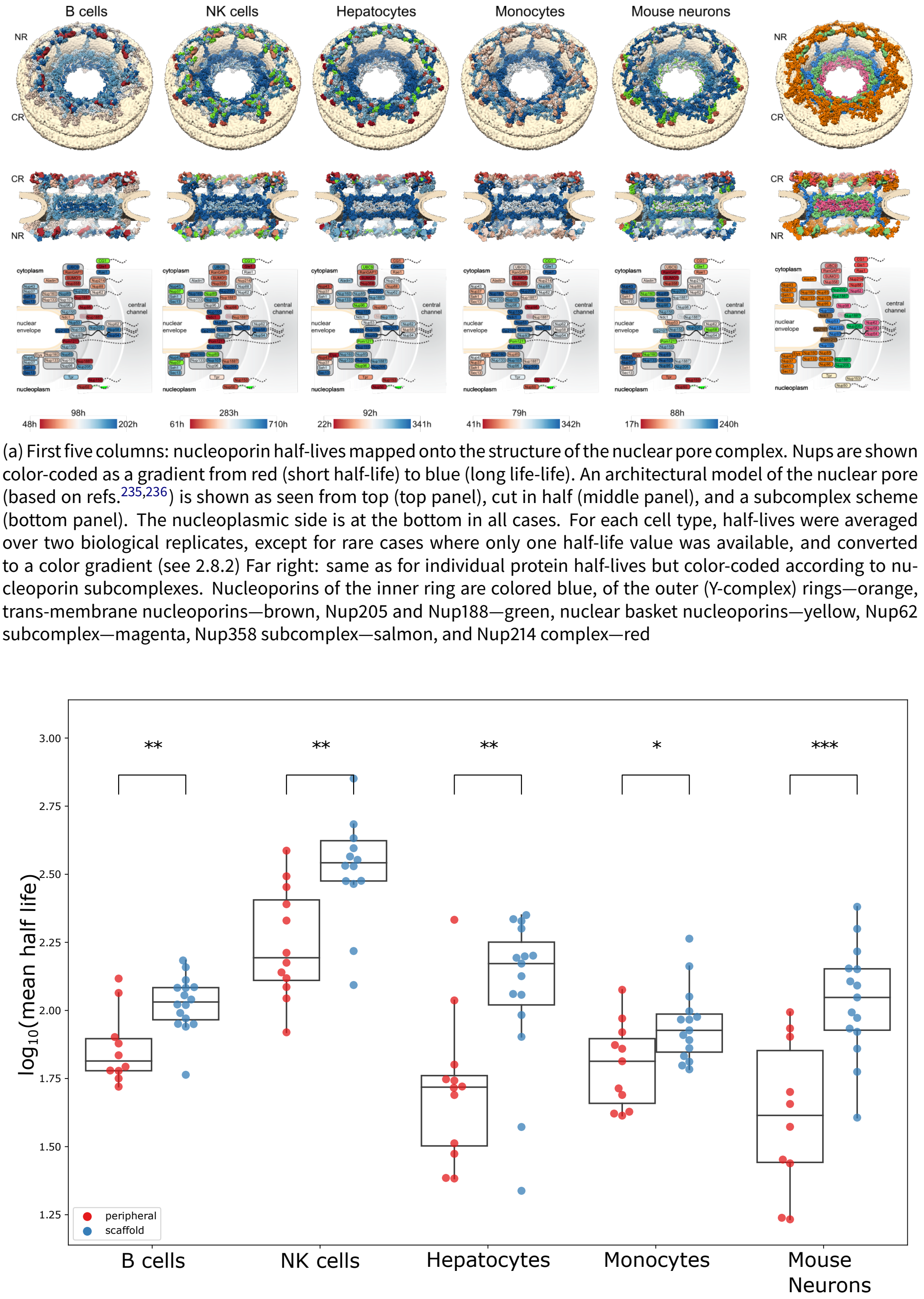


Figure (18): Density distribution of $\log_{10}$ transformed protein half-lives for nucleoporins for each cell type within distribution for all proteins. In all cases the protein turnover for these proteins was towards the middle of the range for all proteins.

4 Discussion

isobarQuant was used to create a catalog of 9699 protein half-lives in five different, non-dividing cell types. This not only provides a high quality resource for the community but also the means to carry out MS1-based quantification. In order to perform highly accurate and precise peptide and protein quantification several new methodologies were developed, implemented and published as part of the isobarQuant software. isobarQuant was shown to perform more accurately than the popular MaxQuant software for determining small fold changes. The first of the new methods leading to this improved performance was the use of an exact model for construction of the isotope envelop, the second was the use of the prior ion ratio for determination of isotope purity. Extensive profiling of the thresholds leading to optimum coverage and accuracy was performed and the effect of excluding zero-imputed fold changes was investigated.

Among all the protein half-lives determined, histones were observed to have very long half-lives, in line with previous *in vivo* work done in rat brain[220]. The determination of the proteasome turnover was in agreement with the values generated in another *in vivo* study[237]. The half-lives recorded for the proteins of the NPC's in all five cell types were, however, much shorter than those recorded in an *in vivo* setting[233], both absolutely and relative to histones. This could, in part, be due to the very different ways in which these two studies were performed (biological context and technical details), but one might expect the accuracy of half-lives measured in non-dividing cells *in vitro* should be more accurate despite the loss of the endogenous setting. Up to now, NPC turnover has been considered a relatively rare event[238], but one of high enough importance to be subject to a surveillance pathway for defective NPC-



(a) First five columns: nucleoporin half-lives mapped onto the structure of the nuclear pore complex. Nups are shown color-coded as a gradient from red (short half-life) to blue (long life-life). An architectural model of the nuclear pore (based on refs.[235,236]) is shown as seen from top (top panel), cut in half (middle panel), and a subcomplex scheme (bottom panel). The nucleoplasmic side is at the bottom in all cases. For each cell type, half-lives were averaged over two biological replicates, except for rare cases where only one half-life value was available, and converted to a color gradient (see 2.8.2) Far right: same as for individual protein half-lives but color-coded according to nucleoporin subcomplexes. Nucleoporins of the inner ring are colored blue, of the outer (Y-complex) rings—orange, trans-membrane nucleoporins—brown, Nup205 and Nup188—green, nuclear basket nucleoporins—yellow, Nup62 subcomplex—magenta, Nup358 subcomplex—salmon, and Nup214 complex—red

(b) Distributions of the reproducibly measured half-lives of the scaffold (blue) and peripheral (red) subunits of the nuclear pore in the different cell types. Differences in the distributions of $\log_{10}$ half-lives were assessed by Wilcoxon rank-sum test (significance levels were encoded as *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$). Center line in box plots is the median, the bounds of the boxes are the 75 and 25% percentiles i.e., the IQR, and the whiskers correspond to the highest or lowest respective value or if the lowest or highest value is an outlier (greater than 1.5 * IQR from the bounds of the boxes) it is exactly 1.5 * IQR

Figure (19): Half-lives and architecture of nucleoporin proteins intermediates[239]. A recent publication by Fornasiero *et al.*[240] measured the turnover *in vivo* of proteins in mouse brain over a longer time period than was performed in this study.The authors showed that the situation *in vivo* is more complicated than in vitro because of the paths that amino acids take in the body of an animal following ingestion. The fate of a pulsed amino-acid depends on the metabolism of the entire proteome, with amino acids being recycled following protein degradation, and re-entering the amino acid pool available for protein neo-synthesis or being excreted from the body. To tackle this they introduced a new mathematical model which takes the incorporation of labeled amino acids into the different pools (soluble i.e. blood, solid i.e. proteins) into account. They made a comparison between *in vitro* determined half-lives of synaptic vesicle proteins and those they acquired *in vivo* to reveal that the distribution of half-lives *in vivo* was much wider and on the whole longer than *in vitro*. The authors attribute the difference to the fact that cultured neurons may still be growing and developing axons and synapses at the time of the measurements. They also state that the half-lives of proteins determined *in vitro* and *in vivo* do correlate, but not extremely well.

In these half-life data we observed that the proteasome and the NPC exhibit structure-dependent turnover. The 26S proteasome is described as having entirely different pathways for its 19S regulatory and 20S core subunits[241], which is mirrored in their uncoupled turnover, indicating that the cap and core must only associate dynamically. An interesting observation is that three beta-subunits of the 20S core proteasome have a turnover much faster than the other members of the 20S core. Since it is difficult to conceive that these subunits dynamically exchange with the assembled 20S core, one could interpret this as representing two different populations of proteasomes that share most, but not all of their subunits and have different turnover times. Protein turnover evidence gathered here relating to the proteasome and the NPC suggest that peripheral complex members have shorter half-lives. This is exemplified by the ADRM1 and PSMD4 members of the 19S regulatory particles whose turnover is much quicker than the other members of the subcomplex and show a highly similar variation in their half-life patterns over all cell types studied here. They are both ubiquitin receptor proteins located at the distal part of the regulatory particle and believed to have been recruited to the complex much later in its evolution[241]. It would be interesting to investigate by looking at transcript information.

The methods developed and described above led to isobarQuant being able to perform better than MaxQuant in the determination of very small fold changes. Shortly after its development another tool for MS1-based quantification was published by Mitchell *et al.*[221] which, in a way very similar to isobarQuant, uses the results of a search engine to go back and interrogate the raw spectral data. They use a Gaussian mixture model to remove interference of individual isotope peaks and perform integrated area under the curve (AUC) XIC quantification after multiple Bi-Gaussian peak fitting over many time points. PyQuant does not, however, provide any QC metric to assess the goodness of fit of isotopic clusters to the measured data and does not seem to take the presence of the prior ion ratio into account.

During the development of isobarQuant and the assessment of optimal running parameters, it was possible to demonstrate that including peptide fold changes imputed as zero negatively affects the accuracy of protein half-life determination. It was shown that protein half-lives are more accurate and precise when such zero fold changes are excluded, provided there is at least one other positive fold change present for the protein.

5 Outlook

The success of isobarQuant in the robust determination of protein half-lives has been described and evidenced above. But, as with any tool there are still a few avenues for potential further development. It might be interesting to investigate the effect of the incorporation of some of the features of other software such as a Gaussian mixture model for pure isotope peak determination or revisit the use of the full AUC for isotope XICs rather than just the peak for use in the determination of the quantification signal, but the impact on the overall protein fold change and in turn protein half-lives is likely to be limited. Adaptations to the filtering of which peptides are used in the calculation of protein fold change could include moving away from the use of a rigid least squares cut offs in favor of weighting the median according to least squares fit might result in more peptides ultimately being used for quantification and thereby improving precision. There might also be some merit in devising a scoring scheme to actually score the goodness of fit, such as that presented in the pyQms[209] software, or increase the weight (influence) of higher scoring or higher-confidence peptide fold changes on the protein fold change.

It is debatable if there is any merit in re-assessing the inclusion of indeterminable peptide fold changes (imputed zeros) by replacing the missing value with the recorded instrument noise value and using that to determine the fold change under the assumption that the true signal is there but simply too low to be recorded by the Mass Spectrometer. This approach would have to be rigorously tested and a thorough assessment of the gains carried out before accepting this method.

One interesting and so far unexplored aspect would be to examine the potential gains of using the peptide fold changes directly in the determination of protein half-lives, rather than distilling this data to a single point and feeding it into a second algorithm. The increase in number of data points would increase the statistical power. This kind of approach is taken by software tools like MSStats[242].

6 Author contribution to project

The majority of data presented in this section was published in reference[181]. Toby Mathieson designed and performed the processing of all experiments in this section using isobarQuant; the laboratory experiments were performed as described in ref.[181]. The author carried out all steps of the data analysis and optimized the filtering parameters using the prior ion ratio and least squares fits. Toby Mathieson designed the layout and procedures of the code and wrote a substantial part of the implementation to perform MS1 quantification. The author investigated the effect of excluding indeterminable fold changes on the numbers of quantifiable protein half-lives for the different datasets and performed all MaxQuant analyses required for the comparison to isobarQuant. The author created all graphics and illustrations in this section (where not separately referenced).

Part IV

Investigation into the effect of TMT labels on peptide fragmentation patterns

1 Introduction

As mentioned above, isobarQuant's approach to storing acquired data alongside the interpreted search results inside the same .hdf5 file enables easy and rapid investigations to be feasible. It is possible to query several tens of .hdf5 files at the same time to enable researchers to look more deeply into any phenomena they might be interested in. This made it the ideal tool for processing datasets to investigate the differences in fragmentation patterns of peptides acquired with a TMT label versus those acquired without any labeling. This is of course possible with other tools such as MaxQuant but often only for one data set at a time. This exploratory study would use several TMT-labeled, offline fractionated datasets acquired with HCD and would compare them to their equivalent unlabeled counterparts treated otherwise experimentally identically.

Higher-energy collisional dissociation[42] (HCD) (described in the main introduction 1.3.4.2) is a beam-type CID in which fragmentation occurs in a dedicated collision cell mitigating the one-third effect associated with traditional CID[243] and is thus able to generate abundant low mass fragments such as immonium ions, the a2, b2 pair, and y1 and y2 ions, which also makes it well suited for use with low-mass, isobaric tagging TMT and iTRAQ quantification. However the spectra derived from HCD have been shown to be more similar to those generated on a triple quadrupole instrument (QqQ-CID) than to conventional CID[244] and statistical characterization of the patterns of HCD spectra compared to CID have shown that HCD tends to generate smaller fragment ions which are typified by many y-ions and a singly-charged, high intensity peak of type b2 which has a high probability of being among the top five most intense in the whole spectrum[245]. The phenomenon of shorter, less abundant, b-ion fragments compared to CID was also observed by Michalski *et al.* who reported extensive y-ion series giving rise to higher peptide sequence coverage with much greater continuous ion series when compared with CID[44].

This apparent b-ion instability and relative increase in y-ion fragments and coverage can be explained if the process of fragmentation in the gas phase is assumed to be similar to CID; such that it follows the 'mobile proton' theory[246,247]. This theory states that the initial site of protonation in peptides is usually at the most basic residues. As soon as the proton is 'activated' by energy from the system it can migrate along the peptide backbone to different, energetically less favorable locations, such as the carbonyl oxygens where it initiates fragmentation at the peptide bond. The initial number of protons in relation to the number of basic residues on a peptide affects the mobility of the ion(s) and can be used to group the peptides into one of three categories (mobile, partially-mobile, non-mobile)[75]. The increased collision energy used for HCD means that mobile protons have greater potential to initiate cleavage at all peptide bonds and as a consequence we see more degradation to a- and lower b-ions in fragmentation pathways described by[248]. Less-mobile protons (such as those in the vicinity of a basic [arginine / lysine / histidine] residue, the first two typically present at the C-terminus of peptides generated by a trypsin digest) are less affected, giving rise to y-ion series fragments similar to CID[245]. The introduction of another, highly basic group at the N-terminus of the peptide in the form of the TMT label and potentially also the C-terminus for lysine-terminating peptides is likely to affect the mobile proton and in turn the peptide fragmentation.

The peaks in an MS/MS spectrum of a TMT-labeled peptide will already differ from non-labeled spectra because of the reporter ions present in the low mass region, but Pichler and colleagues reported that other, non backbone fragments, due to unexpected cleavage within the TMT tag itself, may also be present[249]. This is accompanied by a drop in score for both Mascot and SEQUEST search engines.

The purpose of this study was to investigate the effect of the TMT-tag on the fragmentation of peptides and then apply the results to potentially make improvements to the H-score algorithm published in 2010[165].

2 Materials and implementation

For this investigation two primary datasets collected from an analysis of an *E. coli* digestion standard (Waters Corporation USA). The first was TMT-labeled, where the tryptic peptides were labeled with TMT isobaric tags according to the manufacturer's instructions. The tryptic peptides of the second data set were not labeled. It should be noted at this point that the TMT reagents were the original (six-plex) formulation. Peptides were firstly offline-separated into 16 runs each for 130 minutes using a 75µm ID tip column.

Spectra were acquired on an LTQ-Orbitrap Velos (Thermo Fisher Scientific) coupled to Eskigent nano LC system. The peptides eluted were detected in the LTQ Orbitrap at 30,000 resolution and were subjected to HCD fragmentation with the following instrument settings: Target value FT, 1x10^5 ions; maximum FT fill time 50ms; isolation width, 1.0 Da For the first eight runs a collision energy of 45% was used and for the last eight, 35%. Fragment ions were detected in the Orbitrap at a resolution of 7,500. Raw data were processed entirely using isobarQuant as described in Part II of this report. The pre-Mascot workflow was run and .mgf files were created according to the procedure described in 2.1. All eight runs acquired at the same normalized collision energy (35% or 45%) were merged during the post-Mascot (section 2.2.1) workflow after being searched using Mascot 2.5.1 with the following parameters: 10 ppm precursor mass accuracy, 0.02 Da fragment ion mass accuracy. Variable modifications used were acetylation of protein (N-term), oxidation (M), TMT6plex (N-term); fixed modifications were TMT6plex(K) and carbamidomethylation (C). The maximum number of missed cleavages was set to 3. The instrument type chosen was 'ESI-TRAP' (this setting allows Mascot to match b and y fragment ions and was extended to also include immonium ions) and the enzyme specificity selected was 'Trypsin/P'. Data were searched against Uniprot *E. coli* release November 2017) supplemented with protein sequences of known contaminants (bovine serum albumin and dog, sheep and human keratins). The database contains a total of 12,158 sequences of which 50% are target and 50% are decoy (reversed protein). During the study two further sets of search parameters were used: One with the same modifications as stated previously but using TMT-modifications as 'fixed' and a second one where the decoy sequences were created using a shuffled approach rather than a reverse protein approach.

2.1 Investigation into differences between TMT and non labeled peptides in terms of precursor peptide, Mascot score, peptide length, retention time and fate of triggered MS/MS events

The different tables of the .hdf5 files generated by isobarQuant were queried directly using Python via PyTables in a way similar way to the example given in section 3.1 and plots were created using Matplotlib within a Jupyter notebook. Where necessary, precursor peptides were combined to their Mascot identifications via the MS/MS scan identifier. The origin of peptide identifications (all associated protein accessions and whether these corresponded to target or decoy hits) was parsed from the .dat file during the Mascot parser step of the post-Mascot workflow, internalized into the .hdf5 file and used to determine FDR. When peptides were matched between TMT-labeled and unlabeled datasets, the highest scoring representative within the dataset for a given peptide plus its parent charge was used. Unless otherwise stated, all plots were made using Mascot rank1 peptides filtered to be below a 1 % FDR threshold.

2.1.1 Fragment-ion trend investigation

All rank 1 peptides passing a 1% FDR threshold were extracted from the corresponding .hdf5 files generated during the post-Mascot workflow of isobarQuant. The Mascot-suggested sequence, plus all assigned modifications was used to generate a theoretical MS/MS spectrum comprising of singly charged a-, b-, c- and y-ions against which was compared data from each of the deconvoluted, experimentally-acquired ions from the corresponding deconvoluted spectrum ('deconvions' table in .hdf5 file, 'raw' group). If an ion matched within 20 ppm, the position within the peptide (relative to both the N-and C-terminus) was recorded along with the series (a, b, c or y) and the intensity of the peak, normalized against the most intense peak in the spectrum. Where multiple ions matched within tolerance, the ion with the lowest ppm to the theoretical fragment was selected. Should there still be a tie, the ion was selected according to the hierarchy: y>b>a. The trends for the different series were plotted. This was repeated for all four datasets, considering only the target peptide hits. Appropriate water and ammonia loses from the parent were ignored for this calculation. However, neutral losses for specified modifications were included. In a second step, theoretically generated internal fragment ions and immonium ions (lysine [101.1079], glutamine [101.0715], methionine [104.0534], oxidized methionine [120.0483], histidine [110.0718], pheylalanine [120.0813], arginine [129.114], cysteine (carbamidomethylated) [133.0436], tyrosine [136.0762] and tryptophan [159.0922], taken from http://www.ionsource.com/Card/immon/more.htm) were also matched to the deconvoluted ions stored in the .hdf5 file.

2.1.2 Median peptide coverage ratios for b- and y-ion series

The number of explained cleavage sites (backbone fragments providing evidence for the given peptide bond cleavage of the peptide) was calculated based on the matches made above (2.1.1) for the b- and y-ion series of all peptides. The total proportion of the peptides explained by the given series over the theoretical total, expressed as a median for peptides less than length of 25 residues.

2.1.3 Matrices of cleavage bias for b- and y-ion series

The different amino acid residues either side of all cleavage sites explained in the section above (2.1.1) were recorded, along with the corresponding normalized fragment intensity and series. To gain insight into any cleavage bias, the median normalized intensity of ions explaining each combination of residues was plotted in a matrix for each ion series of each dataset. Each median intensity value was converted to a color within a heat map using the Python Matplotlib package in a Jupyter notebook.

2.1.4 Calculation of complementary pairs

All deconvoluted ions were scanned and each complementary (MH^{2+}– *m/z* of fragment) mass was calculated. The fragment charge is assumed to be +1 as the list is deisotoped and deconvoluted. This calculated, complementary mass was sought within the corresponding table of deconvoluted ions for the given spectrum at a tolerance of 20ppm. Upon finding one or more matches, the fragment with the lowest ppm error was recorded, along with the ion series it belonged to and peptide type (target, decoy or none if not matched). The *m/z* of both members of the pair are then removed from the list to prevent duplicated matches.

2.1.5 Assessment of unassigned ions

A number of potential sources other than instrument ‘noise’ can result in a fragment ion not matching to any of the theoretical backbone fragments. To investigate factors leading to this, any unassigned ion (*m/z* and associated intensities) was recorded if it was within the top 30 most intense ions of a spectrum which had been assigned to a peptide by Mascot. Note that here unassigned means that it was not assignable in the fragment matching step, and is not related to any Mascot fragment allocation. The difference from the parent mass (MH^{2+}) was calculated and recorded. These delta values were then placed into bins of width 0.1 Da and plotted via Matplotlib.

2.1.6 Effect of removal of TMT label-derived ions

The fragmentation of the TMT label at unexpected points within the balancer group can result in the creation of complementary pairs that could be mistaken for b/y pairs. These should be excluded from the complementary pair calculation. If a fragment ion is present in the deconvoluted spectrum and matches within 20ppm to an ion on the list of known TMT-label-derived ions (uncovered during the assessment of unassigned ions above (2.1.5) it was excluded from the calculation of complementary pairs. This filtering step was also included in the creation of .mgf files. These were submitted to Mascot using with the same parameters as above. The post-Mascot workflow was then repeated using the updated search results. Score distributions for the new searches were plotted and a scatter plot pivoting on the highest scoring PSM per peptide for each search was plotted.

2.1.7 Relationship between unassigned complementary fragments and S2I Values

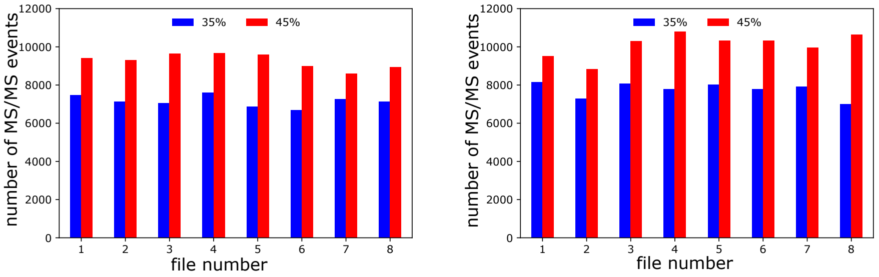
To investigate whether the number of unassigned complementary b/y ions is related to the current proxy for co-eluting peptides, the S2I value, the counts of unassigned complementary pairs for each spectrum (see section 2.1.5), and excluding those pairs derived from TMT-balancer fragmentation) were plotted as a function of the calculated S2I of the peptide and summarized as a heatmap with red denoting the highest counts, blue the lowest.

2.1.8 Incorporation of contiguous explained sites into H-score

Cleavage sites (where a backbone fragmentation occurred and the corresponding b- or y-ion had been identified) were calculated for all Mascot rank 1 hits (both target and decoy) as described for the fragment ion trend investigation (2.1.1). These values were then used to calculate the maximum number of adjacent cleavage sites for both b- and y-ion ladders and combined into a single value. This value is the contiguous explained sites and was recorded. It was divided by the maximum possible total cleavage sites (the peptide length minus one) to normalize for different peptide lengths. The H-score then proceeded as follows: if all possible cleavage sites were explained an additional 3 points were awarded to the total number of explained cleavage sites (as for the published H-score). If the ratio was greater than 0.1 then an additional two points were awarded.

3 Results

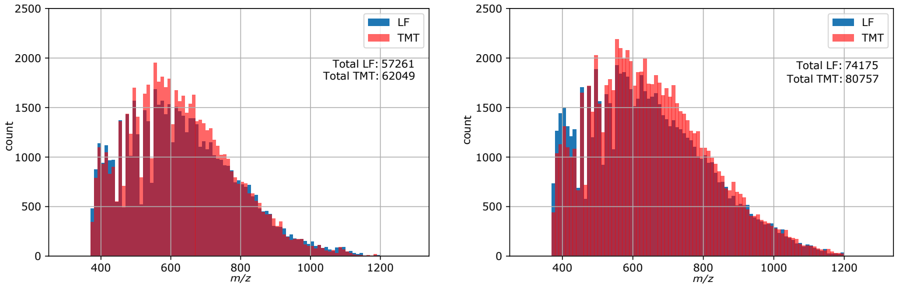
Data were processed in the same way as described in the Methods section using the isobarQuant pre- and post-Mascot workflows. A total of four different conditions are compared; + / - TMT label at two different collision energies. The resulting raw and interpreted data stored in .hdf5 files by isobarQuant are readily extractable and can be



(a)

(b)

Figure (1): The number of MS/MS events triggered per file for TMT labeled and unlabeled peptides. Figure (a) displays the unlabeled peptides and figure (b) those labeled with TMT reagent. The higher collision energy yields a higher number of MS/MS spectra irrespective of its label status. The count of MS/MS events does not differ significantly between the two label states nor does it between the individual files.



(a)

(b)

Figure (2): Distribution of acquired precursor ions triggering MS/MS events for combined datasets at 35% and 45% nCE. Precursor data were extracted directly from isobarQuant-generated .hdf5 files. A global increase in precursor size can be observed for TMT labeled peptides (red / pink) compared to those lacking the TMT group (blue). It is also possible to observe the instrument's fixed lower-mass cut off at around 375 Th.

easily combined to create the following summaries and facilitate this investigation.

Focus is first on the uninterpreted MS/MS data. Figure 1 displays the frequency of triggered MS/MS events: there is no significant difference within or between the individual files of labeled and unlabeled datasets. Whilst there is no significant increase in the frequency of labeled precursor masses, the total count of MS/MS events is around 8% higher for TMT labeled peptides at both collision energies (Fig. 2). In the same figure (Fig. 2) we see that the number of MS/MS spectra acquired is 23% higher at 45% nCE for combined labeled and unlabeled data [154,932 and 119,310 respectively] and that the addition of the TMT tag shifts the precursor masses (*m/z*'s) to the right.

There is an increase in the mean number of acquired fragment ions per spectrum for peptides labeled with TMT at both collision energies, with the higher collision energy yielding more fragments. The reduction in mean number of ions per spectrum resulting from deconvolution (isobarQuant removes TMT reporter ions at this stage, see chapter 2.1.6) is around 12% for unlabeled samples and between 26 and 29% for

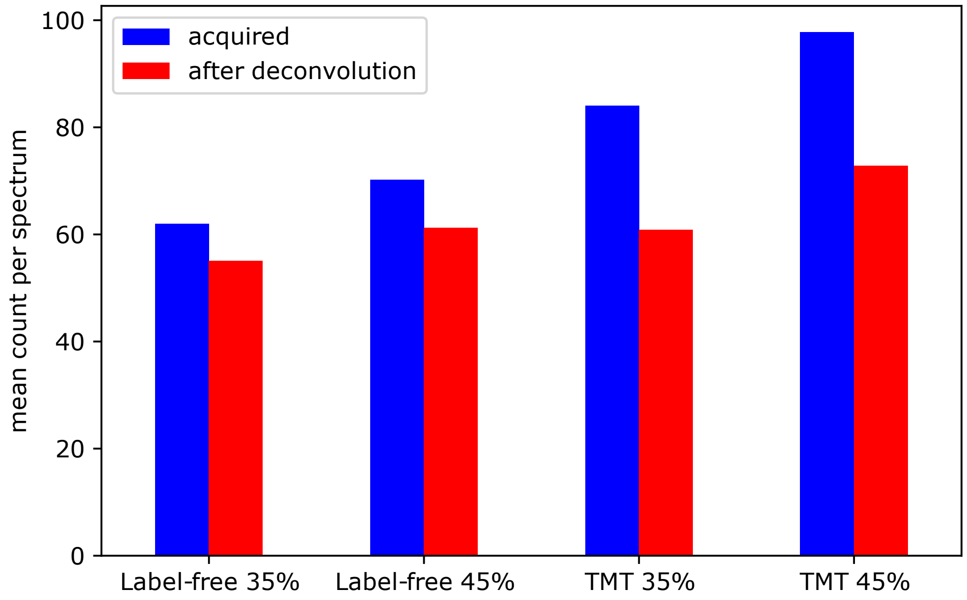


Figure (3): Mean number of fragment ion counts per MS/MS spectrum for the four datasets. MS/MS data was extracted directly from the isobarQuant-generated .hdf5 files. The mean number of fragment ions per spectrum is higher for TMT labeled peptides (between 80 and 100 compared to 60 and 70 for unlabeled). Following deconvolution, the counts do not significantly differ between the label states. During the deconvolution stage of isobarQuant processing the reporter ions are automatically removed from the spectra

TMT, 35% nCE and 45% nCE, respectively (Fig. 3). This increase is greater than can be attributable to reporter ion removal alone. Overall, these results show that the addition of the TMT tag results in a small increase in the number of triggered MS/MS events, a general increase in precursor *m/z* and an increase in the number of fragment ions per spectrum.

In order to gain further insight into the effect of the addition of the TMT group the investigation moved to looking at differences in search results, matched PSMs, and peptide fragmentation. The isobarQuant-generated .mgf files were submitted to Mascot for searching and the resultant peptide and fragment hits stored in the .hdf5 files were assessed.

The distribution of Mascot peptide target hits is loosely in line with the numbers of the acquired spectra at the given *m/z*, with TMT labeled spectra yielding a larger proportion of targets than unlabeled spectra. The maximum proportion of MS/MS spectra explained by a target sequence, Fig. 4, yellow bars, is 25% (15,733 / 62,049) for TMT, 35% nCE and the minimum (15% 10,831 / 74,175) is label-free, 45% nCE. The increased number of Mascot-assigned target peptides for both TMT datasets is substantially larger (>20%) than can be accounted for by the apparent gain in triggered MS/MS events. There is, however, also an increase in the number of decoy peptides of around two-thirds. For 35% nCE the number increases from 1,143 to 3,426 and for 45% nCE from 933 to 2,373 (an increase of more than two thirds of the TMT total). This increase in decoy peptides (Fig. 4, green bars) results in a higher global false discovery rate, and at 35% nCE leads a lower number of target peptides passing the 1% FDR

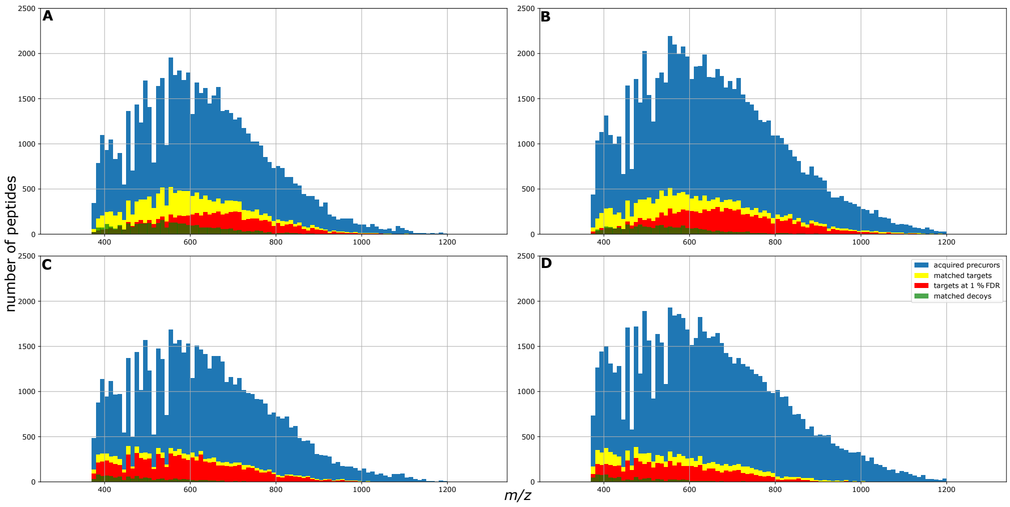


Figure (4): Histograms describing the fate of triggered MS/MS events for all four datasets. The upper panel (A & B) shows TMT labeled data and the lower (C & D) depicts unlabeled. Plots on the left are acquired at 35% nCE and the right, 45% nCE. MS/MS spectra unassigned by Mascot are shown in blue, spectra matching to target peptides are shown in yellow and those matched with a FDR lower than 1 % are overlaid in red. All matches to decoy peptides are shown in green.

filter (from 16% [9,406 / 57,261] to 13% [7,933 / 62,049]) (Fig. 4, red bars). However, despite the increase in decoy peptides, this decrease is not observed for the 45% nCE sample where the number of targets passing the 1% FDR cut off is 9% [6,784 / 74,175] in the non labeled sample compared to 13% in the labeled sample [10,066 / 80,757]. To rule out that the possibility that any of these effects could be related to the method of decoy peptide creation (here using the reverse protein approach) all TMT searches were repeated against a shuffled-peptide database. The results were almost identical with no substantial changes in numbers of decoys or frequency of target peptides matched.

The change in score and peptide length distributions following TMT labeling is shown in figure 5: the addition of the TMT label results in narrower score distributions with fewer peptides at either extreme. At 35% nCE, the average Mascot score for a peptide with a TMT label is 45 compared to 51 for unlabeled, with the maximum score for unlabeled peptides being 163 compared to 131 for TMT. The mean Mascot score decreases to 35 and 36 for TMT and unlabeled peptides, respectively, acquired at 45% nCE, with much lower maximum scores of 95 for TMT and 116 for unlabeled. The number of peptides with a score >50 at 35% nCE is 2,550 compared to 4,569 for the unlabeled data (1,094 and 1,366 for 45% nCE, TMT and unlabeled respectively). The numbers of unfiltered target peptides for labeled and unlabeled data at 35% nCE (Fig. 5A, dashed line) show similar trends but the counts associated with the labeled peptides is around double that of the unlabeled peptides. This pattern is repeated at the higher collision energy but to a much lesser extent (Fig. 5B)

The lower panel highlights that target TMT-bound peptides are shorter compared to those without a tag at both collision energies, with an average length of 9 (10, < 1% FDR) and 11 (11, <1% FDR) at both 35% nCE and 45% nCE respectively. The trends

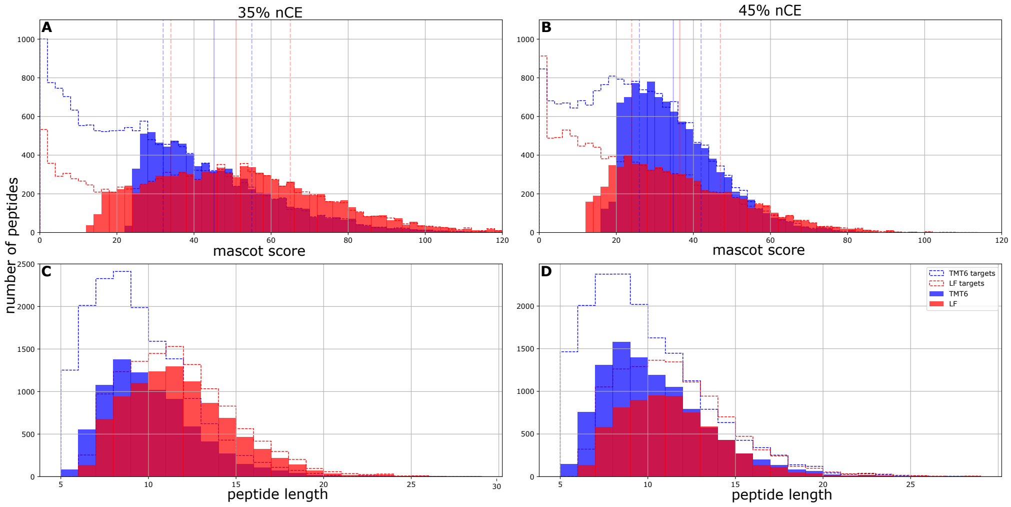


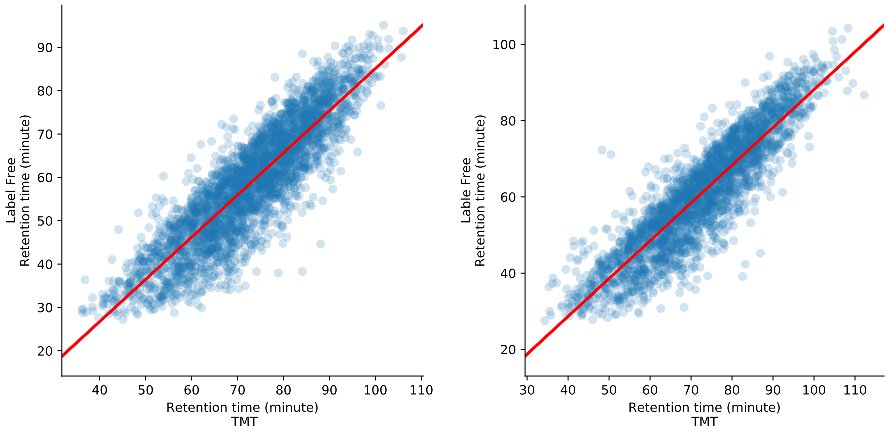
Figure (5): Distribution of score and peptide length for target peptides from all datasets. The solid bars represent those peptides passing the 1% FDR criteria, the dashed lines outline for all. The upper panel (A&B) displays the score distribution of TMT labeled peptides (blue) and unlabeled (red). The solid line represents the mean mascot score for the high-confidence data with the dashed lines indicating the upper 25% and lower 75% bounds. On average, label-free peptides achieve higher scores than those labeled with TMT. The spread is wider for unlabeled data. The lower panel (C&D) summarizes the length of identified peptides for TMT labeled peptides (blue) and unlabeled (red). The left side is the lower collision energy (35% nCE) and the right side is 45% nCE) There are more target peptides identified for TMT labeled data than for unlabeled and on average peptides are shorter (i.e. consist of fewer amino acids).

are similar for both high quality and non-filtered target peptides, again with a higher number of target peptides identified with a TMT label than without one.

The increase in molecular mass and physico-chemical property changes brought about by the presence of the 229 Da aromatic ring will have an effect on the behavior of the peptides in the LC column. To investigate this, isobarQuant-generated .hdf5 files were queried and the RT of peptides passing the 1% FDR filter, which were observed in both the labeled and label-free experiments were plotted against each other. A linear fit was made to the data points to assess any shift in retention time between the two methods. The presence of the TMT tag increases the retention time by between 11 and 12 minutes (Fig. 6) with no significant difference between the experiments at different collision energies (which underwent LC-separation on the same column).

3.1 Effect of tag on Mascot scoring

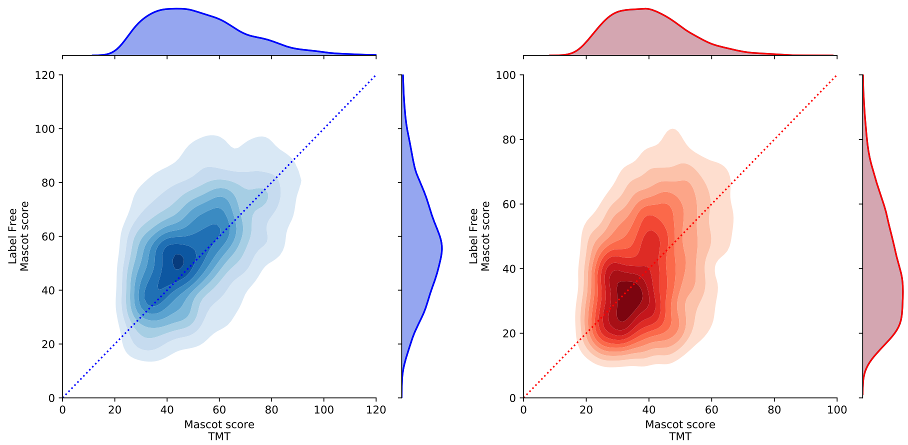
Extending the investigation of the effect of the TMT tag on Mascot score, a subset of FDR -filtered (< 1%) peptides occurring in both labeled and unlabeled sets (35% and 45% nCE respectively) were compared. The trend in peptide scores reveals that at the lower energy (35% nCE) a peptide generally scores better when not labeled with TMT, Fig. 7a. However, for 45% nCE this trend is much less pronounced (Fig. 7b). The spread of the data is quite large at both collision energies. The mean score delta between TMT-labeled peptides and non-labeled is -5.1 for 35% nCE (with a σ of 20) and –0.7 for 45% nCE (with a σ of 17) suggesting that, while a TMT tag is likely to lead to a reduction in Mascot score, this is far from a universal phenomenon.



(a) 35% nCE

(b) 45% nCE .

Figure (6): Retention time alignments between (the best scoring) identical peptide sequences found in TMT and label-free samples. The slope of the linear model fit shows that peptides elute between 11 and 12 minutes later when they are labeled with TMT. (35%: 10.98', 45%: 12.15')



(a) 35% nCE

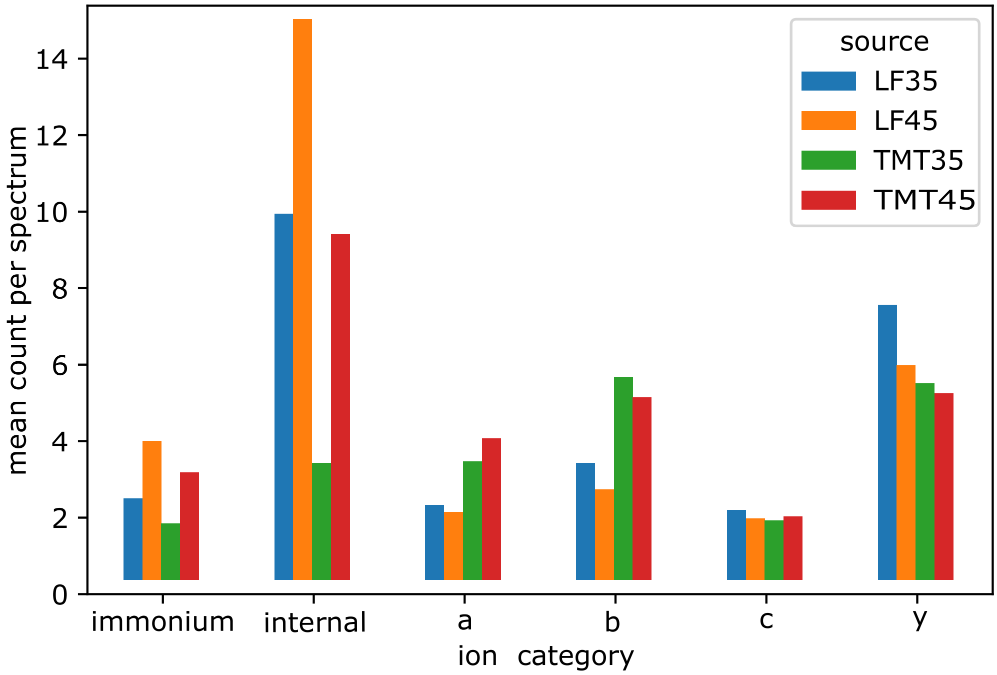
(b) 45% nCE

Figure (7): Two-dimensional density plots of Mascot score overlap between identical peptide sequences labeled or unlabeled with TMT. Non-labeled peptides generally obtain better scores in the 35% nCE data set but not in all cases. The effect was much less prominent in the scores of the 45% nCE.

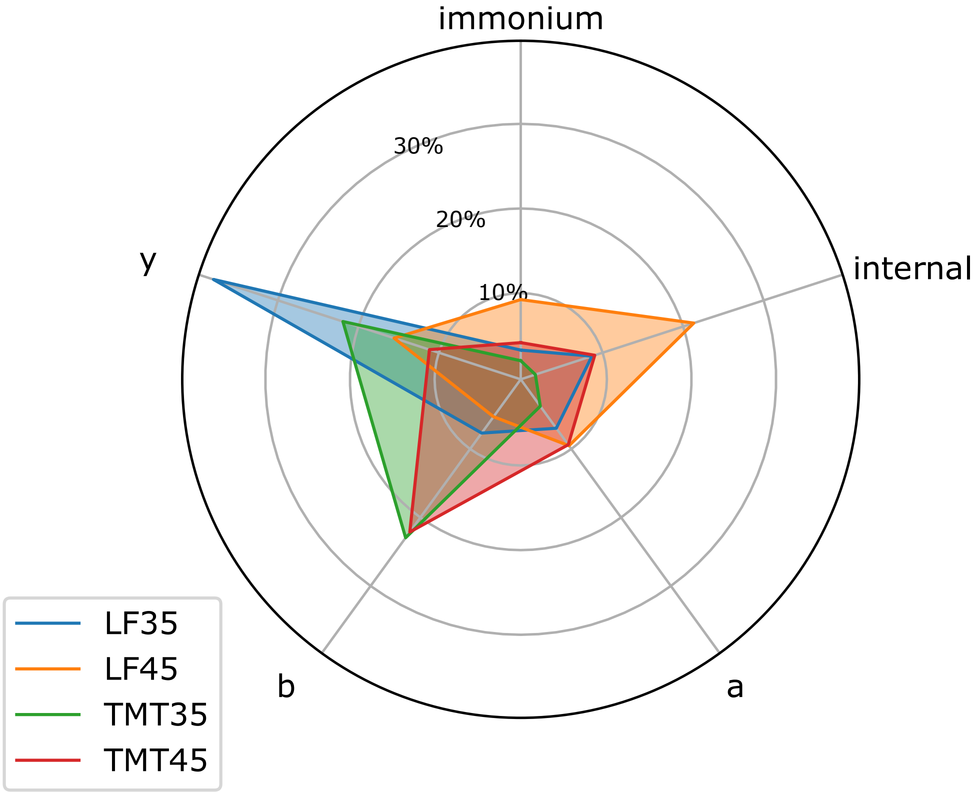
With the investigation into the general trends of peptide differences between the four samples complete, the .hdf5 outputs from isobarQuant were used to review the different fragment ion series present within the Mascot-assigned PSMs passing the 1% FDR threshold. Theoretical masses for four series of backbone ion fragments likely to occur in HCD fragmentation (a, b, y and, to a lesser extent, c) and all possible internal ion fragments were calculated. The fragments in these five theoretical fragment categories were based on the peptide sequence assigned by Mascot. The sixth and final fragment ion category corresponded to the masses of ten known immonium ions likely to be visible in HCD fragmentation. The deconvoluted and deisotoped peaks of the MS2 spectrum stored in the .hdf5 file were then matched to all corresponding generated fragments within a tolerance of 20ppm. In cases where one measured ion matched to more than one theoretical fragment within tolerance the lowest delta match was used. No intensity cut off was applied and no false discovery rate was determined. The intensity of each measured ion was normalized to the most intense ion in the spectrum and recorded along with the intensity as a proportion of the TIC. For each of the six fragment categories the mean count of corresponding fragments per spectrum was tallied and the normalized and proportional fragment intensities were summed.

These data show that use of higher collision energies results in just under double the mean number of immonium ions per spectrum and the presence of a TMT label reduces this mean number by about 30% (Fig. 8a). When translated to the proportion of signal intensity per spectrum, less than 10% of total signal was generated by immonium ions (Fig. 8b) with the highest mean proportion (9.2% of total signal) coming from the label free data acquired at 45% nCE. The lowest mean proportion was seen with TMT labeled peptides acquired at 35% nCE with 2% of the total signal. The trend seen for internal fragment ions is similar to that for immonium ions. The higher collision energy, on average, yields the highest number of internal fragments per spectrum 15 (unlabeled) and 9 (labeled) with six fewer fragments in each case for the lower collision energy (9, unlabeled, and 3 labeled). Therefore the presence of a TMT tag results in about one third fewer internal fragments (Fig. 3). The mean proportion of total spectrum signal is similar to the counts per spectrum. TMT labeled data acquired at 35% nCE shows only around 2% of total signal attributed to internal fragments. This number rises to just below 10% for the label-free, 35% nCE and for TMT labeled, 45% nCE. The greatest proportion of signal derived from internal fragments is found in the unlabeled sample acquired with the highest collision energy and represents over 20% of the total spectrum intensity (Fig. 8b). It is interesting that the internal ion peak counts per spectrum represent the highest mean number of fragments of all six fragment categories, with the exception of TMT labeled, 35% nCE peptides. Of all the back-bone fragments, the y-ions are most prevalent in every data set with an average count of between five and eight peaks per spectrum (Fig. 8a). The proportion of total intensity is always greater than 10%, with the highest value (just under 40%) observed with the unlabeled, 45% nCE data set. TMT labeled, 45% nCE shows the lowest total signal when expressed as a percentage of the TIC. The greatest difference between label free and TMT samples is seen in the similar trends of the a- and b-ions. There are about twice as many a- and b-ion fragment counts per spectrum in TMT data than there are in unlabeled peptides; this trend is amplified when looking at the proportion of total signal coming from b-ions (> 20%) compared to only 5% for non-labeled samples. A-ions never represent more than 10% of the total signal in any data set and c-ion fragment counts are similar in all datasets and, as would be expected, account for less than 2% of the total signal and, as such, are not shown in Fig. 8b.

The most abundant internal fragments originate from the highest collision energies and are represented by the shortest versions, with di-peptides (fragments consisting of two amino-acids) accounting for more than three times the signal of the



(a) Mean matched fragment ion counts per spectrum for the six fragment ion categories. The different colors represent the different datasets.



(b) Proportion of the TIC of a spectrum represented by each ion type, as arranged radially around the center. Each data set is represented by a different color and the average proportion per spectrum is given on the by the node of each spoke.

Figure (8)

next most abundant set of internal fragments (tri-peptides) in all cases (Fig. 9a). The difference in abundance between the most (label free, 45% nCE) and least abundant (TMT, 35% nCE) is around 60-fold for the di-peptides. The presence of the TMT tag reduces the number of internal fragments for all lengths, and the abundance of internal fragments above five or six amino acid residues is extremely low and as seen previously, label-free peptides yield more internal fragments.

Figure 9b shows that the most abundant immonium ions originate from histidine, phenylalanine, tyrosine and arginine and that the general trend observed for immonium ions described above (Fig. 8a) is reflected in the summed normalized intensity for each of the immonium ions; with two notable exceptions. Firstly, the abundance of the histidine immonium ion is greater in samples with the addition of the TMT label (or, at least, is not substantially reduced) and secondly the TMT labeled samples yield no, or very few, arginine immonium ions. Once again the higher collision energies yield a higher abundance of immonium ions.

The most striking difference in the summed normalized fragment ion intensities of doubly charged TMT and label free samples is the abundance of b- and a-ions. Rather than displaying the classic pattern of weaker a- and b-ion series dominated by highly abundant b2 and a2 ions which is usually associated with HCD fragmentation[44], a much more even distribution is evident (Fig. 10). At the lower collision energy, the abundance of b- and y-ions is similar and higher than that of the a-ions. The b-ions dominate the higher energy TMT labeled sample and overall there is higher intensity observed for shorter fragment ions. The a- and y-ions follow a similar trend after the y/a-1 ion. The trend in y-ions is similar for both labeled and non-labeled data fragmented at 45% nCE; but for 35% nCE the classic CID / HCD pattern for a greater prevalence of longer y-ions is missing in the TMT sample. The c-ions are present at very low abundance in all samples and have a slight increase in the non-labeled data. To assess whether or not the increase in b-ions is linked to the sequestration of a proton at the N-terminal TMT group, the fragmentation patterns of +2 lysine and arginine terminating peptides were compared. Figure 11 shows that the b-ion intensity for arginine-terminating peptides is high and well distributed across many positions on the peptide but the y-ion abundance is much lower, which would be expected if the TMT group had greater affinity for the proton(s) than the terminal arginine. The distribution of b- and y-ions on the lysine terminating peptides is, however, much more similar indicating that the b-ion stabilization is related to the N-terminal TMT tag.

3.2 Peptide coverage by different ion types

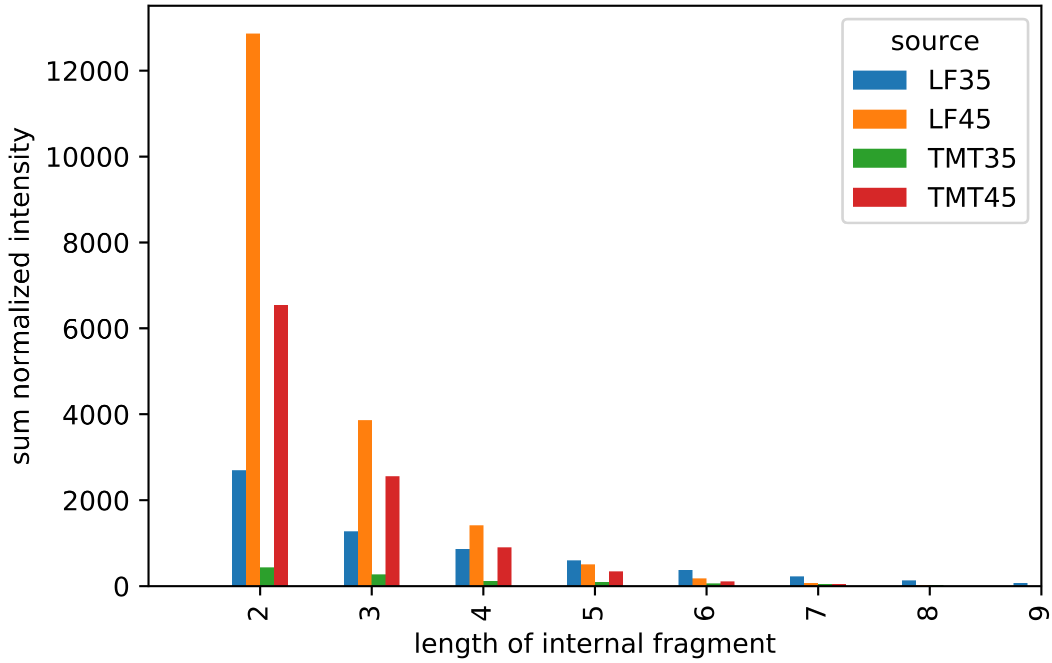
This behavior translates into a higher proportion of peptides being covered by both (b and y) ion types (b-ion coverage remains greater than 50% for peptides up to lengths of 11 or 15 (35% nCE / 45% nCE, respectively), compared to unlabeled peptides, where the b-ion coverage remains below 50% at all lengths of peptide, as depicted in figure 12. The pattern of coverage for the decoy peptides is similar to the target peptides but covers a much lower proportion of the total peptide at all different peptide lengths. This means that decoy peptides show a much higher coverage for b-ions in the labeled samples than the unlabeled and could be one explanation for the increased number seen in figure 4.

The change in fragmentation patterns brought about by the TMT tag was further investigated by mapping the local environment surrounding the b and y backbone fragment ions. The amino-acids pairs between which the peptide fragmentation (cleavage) occurred were recorded and plotted in figure 14. The residue to the N-terminal side (left) of the cleavage corresponds to the amino acid given on the *y*-axis and the residue on the C-terminal (right) side of the cleavage is given by the residue on the *x*-axis. Looking first at the matrix representing b-ions, the increase in intensities of b-ions for the TMT samples (fig. 13) is again evidenced by the overall increase in heat. For the 35% nCE sample, the proline effect is more pronounced than in all other samples: the cleavage **x|P** (’x’ denoting any amino acid) has the highest intensity and, correspondingly, those with proline at the N-terminal position are present at low abundance and are lower for TMT-labeled samples than for the label free ones. Cleavages involving methionine at both the N- and C-terminal are rarer for TMT than for unlabled sample, but this effect is much less pronounced with 45% nCE, where the C-terminal methionine does not seem affected at all. The intensities of the acidic residues at the position N-terminal to the cleavage site show increased abundance in TMT-labeled peptides, with aspartic acid being among the most intense N-terminal residues in the matrix. This is mirrored with a corresponding decrease in C-terminal acidic residues. The intensity of histidine residues C-terminal (**x|H**) to the cleavage site flips from being low intensity in label-free peptides to being quite intense for TMT samples at both lower and higher nCE samples. The differences described here seem to be greater than can solely be attributed to the overall increase in b-ions.

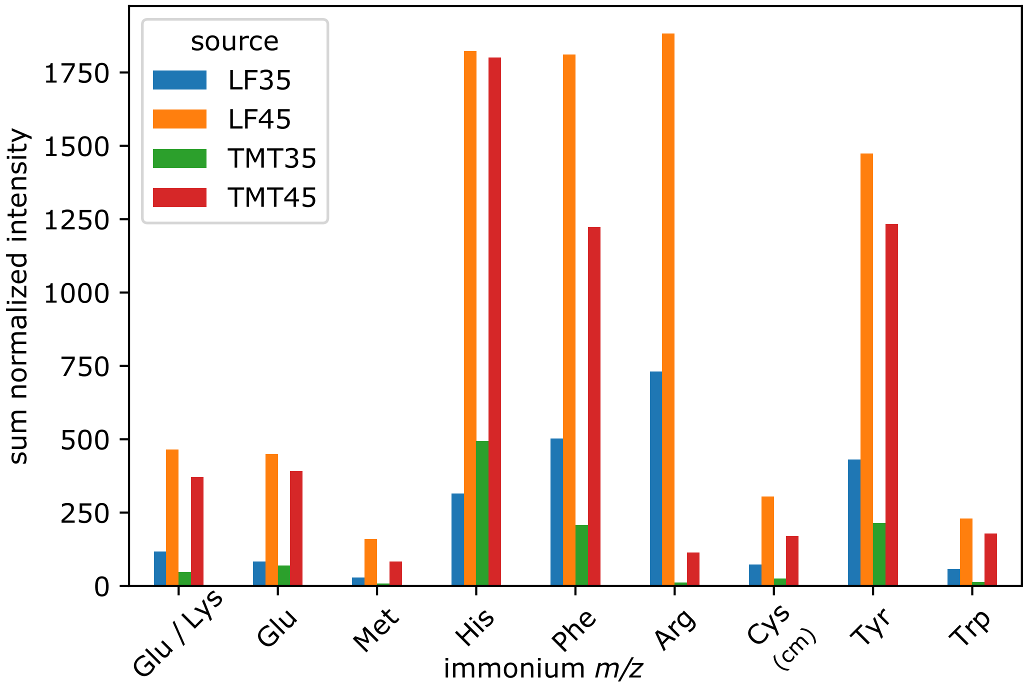
The difference between the two datasets for the y-ions is much less pronounced than for the b-ions, with an overall decrease for ion intensity for TMT-labeled peptides (again due to the relative increase in intensity of the b-ions). The methionine residues are largely absent for the lower collision energy TMT samples, as observed for the b-ions. Apart from that, and the stronger decrease in **P|x** cleavages (proline flanked C-terminally by any amino acid) also observed with the b-ions, there are not as many differences between the labeled and unlabeled datasets for the y-ions.

3.2.1 Investigation into complementary b/y ions

The increased b-ion stability and their increased distribution in the TMT labeled peptides led me to question whether this could be beneficial in the validation of PSM



(a) Sum of normalized fragment intensities for internal fragment ions identified in the different datasets



(b) Sum of normalized fragment intensities for immonium ions identified in the different datasets. Lower sum intensity for TMT peps (less breakdown of b ions) with the exception of histidine

Figure (9)

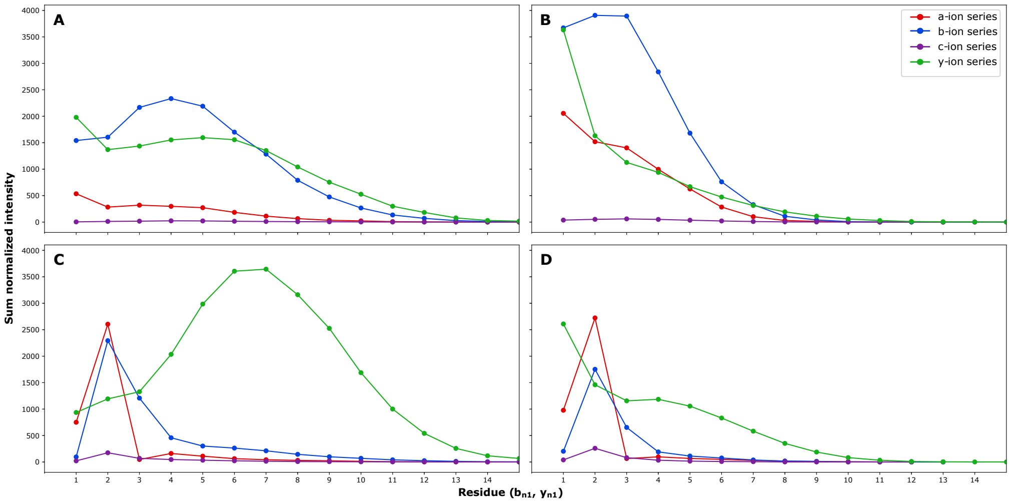


Figure (10): Sum of normalized intensities for backbone fragment ions for amino acids at their position within the peptide relative to N- or C-terminus. Peptides precursors were all of charge state +2 . The upper panel displays the values for TMT labeled peptides (35% nCE [A] and 45% nCE [B]) and the lower panel the values for the unlabeled peptides (35% nCE [C] and 45% nCE [D]).

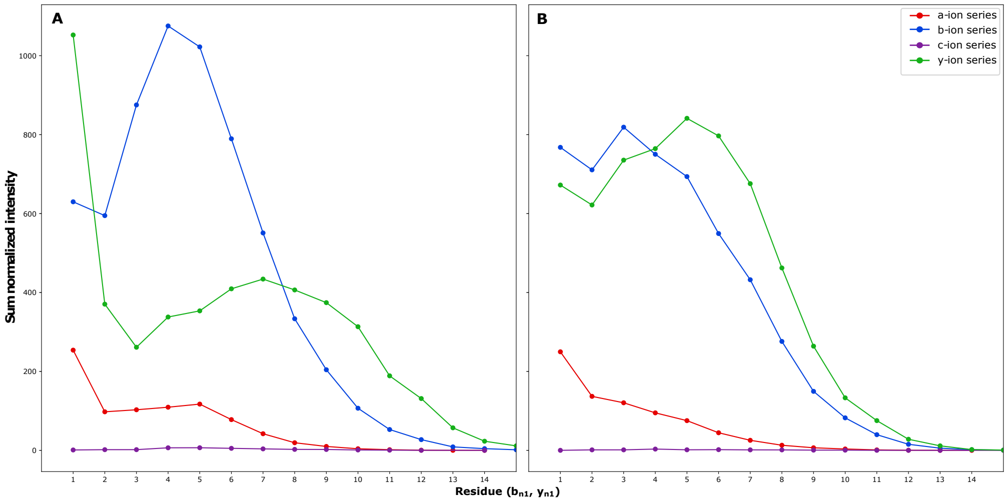


Figure (11): Sum of normalized intensities for backbone fragment ions at their position within the peptide relative to N- or C-terminus. Displayed are only +2, TMT-labeled peptides acquired at 35% nCE which terminate with either arginine (Panel A) or lysine (Panel B). The b-ions remain abundant for arginine terminating peptides while the y-ions are much less intense. The spread of abundances is much more even for lysine terminating peptides

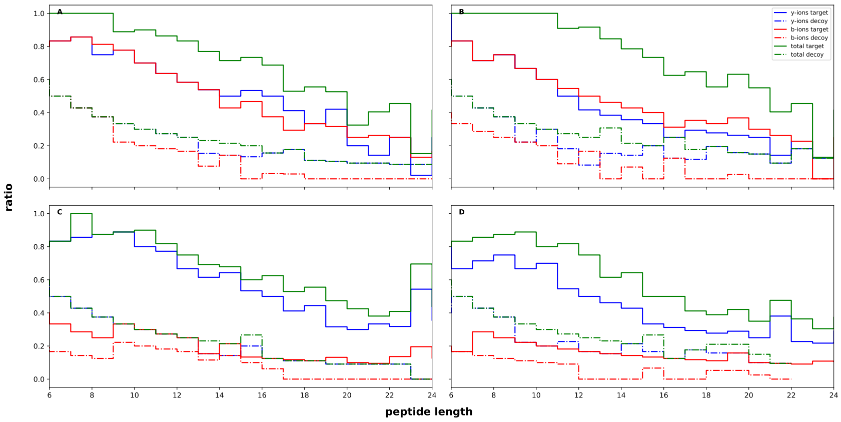


Figure (12): Total peptide coverage ratios provided by different fragment ion types. The upper panel displays the values for TMT labeled peptides (35% nCE [A] and 45% nCE [B]) and the lower panel the values for the unlabeled peptides (35% nCE [C] and 45% nCE [D]). In label-free samples at both collision energies the b-ion peptide coverage remains below 50% for all lengths of peptide compared to the TMT samples where it remains above 50% up to length 16 (35% nCE) and 11 (45% nCE). The decoy peptides' coverage in terms of b- and y-ions reflects that of the target peptides in all cases but with a much lower coverage.

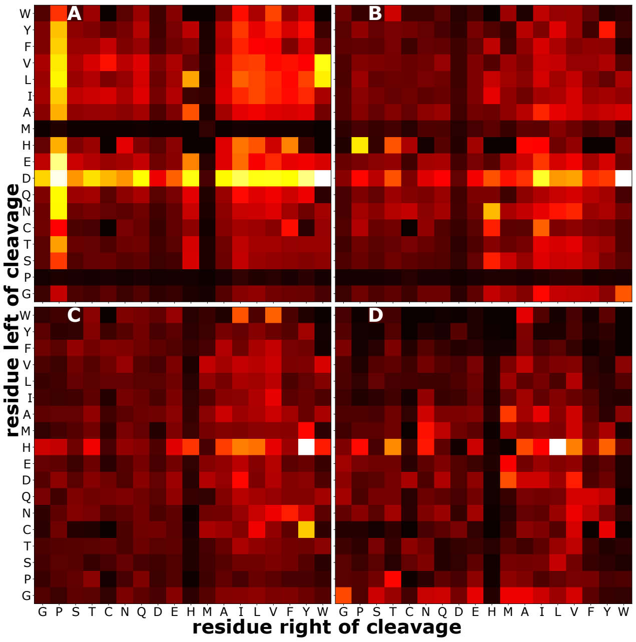


Figure (13): Median of normalized b-ion intensities. The upper panel displays the values for TMT labeled peptides (35% nCE [A] and 45% nCE [B]) and the lower panel the values for the unlabeled peptides (35% nCE [C] and 45% nCE [D]). The increased intensity for the b-ions in TMT-labeled peptides is seen in the greater overall heat. The proline effect seems more pronounced in the labeled sample at 35% nCE and cleavages involving methionine appear rarer for TMT than for the unlabeled peptides. The intensities of the acidic residues at the position N-terminal to the cleavage site show increased abundance in TMT-labeled peptides, with aspartic acid being among the most intense N-terminal residues.

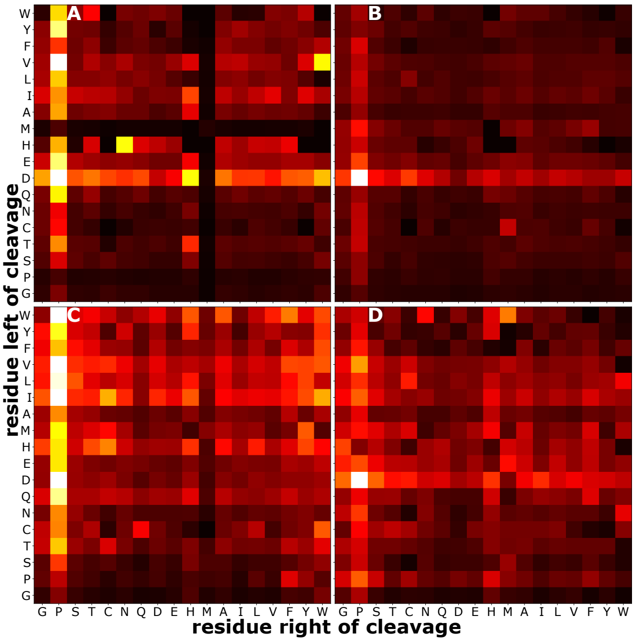
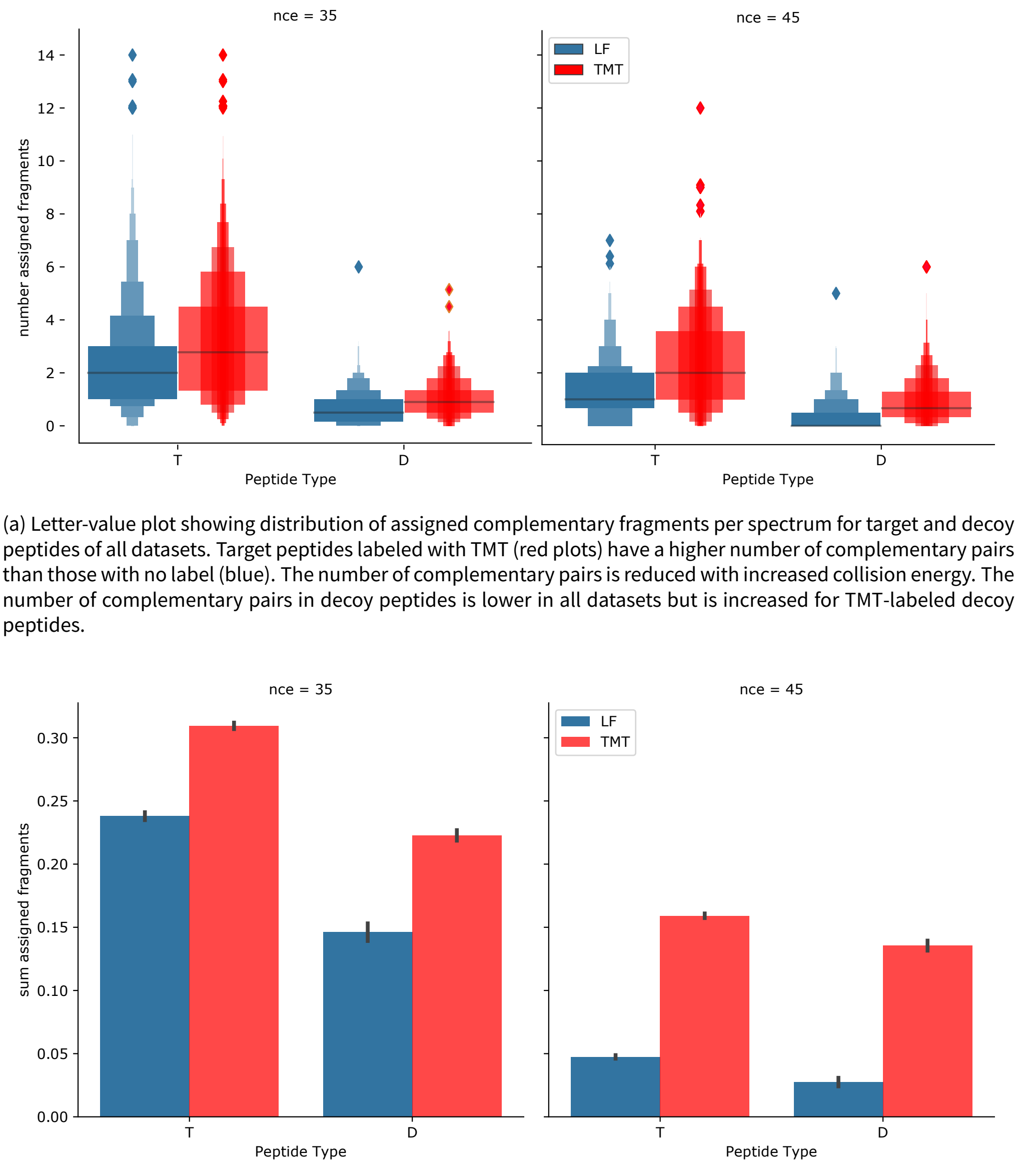


Figure (14): Median of normalized y-ion intensities. The upper panel displays the values for TMT labeled peptides (35% nCE [A] and 45% nCE [B]) and the lower panel the values for the unlabeled peptides (35% nCE [C] and 45% nCE [D]). Differences between the labeled and unlabeled sets are less apparent for y-ions than for b-ions. Overall heat is reduced for the y-ions for labeled data and methionine residues are largely absent for the 35% nCE set.

matches in a way similar to the method developed by Nielsen *et al.*[250], where complementary b- and y-ion pairs represent strong evidence for the existence of a given cleavage site. To investigate the co-occurrence of pairs of b- and y-ions whose summed deconvoluted mass equals that of the precursor (within the given tolerance), each MS/MS spectrum linked to a Mascot suggested peptide was interrogated and the complementary fragment for each of its deconvoluted ions was sought among the remaining ions. Each match within 20ppm tolerance was recorded. The type of match (whether the matched ion corresponded to a member of the theoretical back-bone fragment series or not) was also noted. In order to gauge the results against known false positives, decoy peptides were included in the assessment.

The increase in b-ions for TMT labeled peptides does indeed lead to an increase in the the mean number of complementary b/y pairs per MS/MS spectrum compared to non-labeled (Fig. 15a). This trend is reflected in the proportion of pairs mapping onto assigned backbone fragments with a significantly higher number being identified in TMT spectra than in the unlabeled. For both the unlabeled and the TMT samples the numbers of pairs is highest at the lower collision energy. For all decoy peptides, the overall count of assigned b/y pairs per spectrum is lower than for the targets. This trend is repeated in the mean proportion of TIC for each MS/MS spectrum. The TMT labeled data at 35% nCE has one third of the TIC for assigned complementary b/y pairs, dropping to 15% of signal for 45% nCE (Fig. 15b). This value is lower for label free data at both collision energies; substantially so for 45%. The assigned b/y pair signal was lower for decoy peptides in all cases, but the ratio between target and decoy peptides was the same in all datasets.

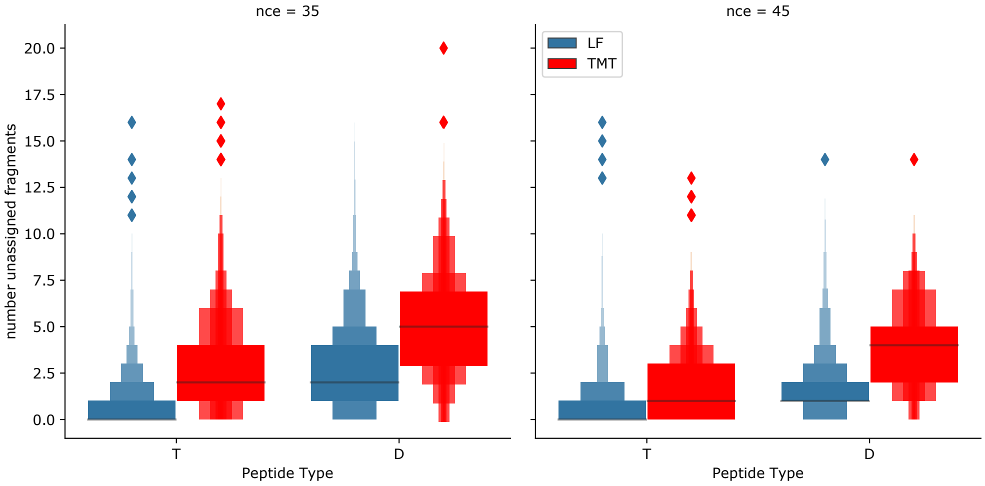
Examining unassigned complementary b/y pairs (fragments not attributable to back bone cleavage of the given precursor but summing to its neutral mass) we see



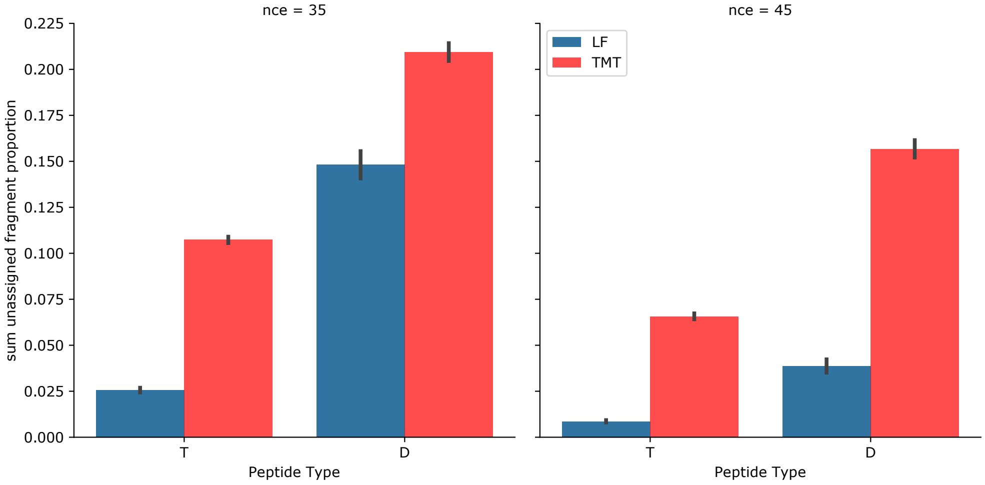
(a) Letter-value plot showing distribution of assigned complementary fragments per spectrum for target and decoy peptides of all datasets. Target peptides labeled with TMT (red plots) have a higher number of complementary pairs than those with no label (blue). The number of complementary pairs is reduced with increased collision energy. The number of complementary pairs in decoy peptides is lower in all datasets but is increased for TMT-labeled decoy peptides.

(b) Histogram of proportion of total TIC attributed to complementary pairs for the different datasets. Bars in red correspond to the summed proportion of total ion intensity for TMT-label complementary fragments, bars in blue are the unlabeled samples. On the left side data are shown for the lower (35% nCE) collision energy and on the right side for the higher (45% nCE).

Figure (15): Overview of amount of fragment ion signal associated with complementary b/y pairs. In terms of both number of fragments and proportion of total signal, the TMT label increases the amount of complementary b/y pairs. This phenomenon is more prevalent at the lower collision energy.



(a) Letter-value plot of distribution of unassigned b/y pairs for target and decoy peptides of both datasets



(b) Histogram of proportion of TIC of each unassigned complementary pair for the different datasets. For TMT labeled peptides, a much greater proportion of the total ion intensity is associated with unassigned complementary b/y pairs (those fragments which are not attributable to backbone cleavages but whose *m/z*'s sum to the mass of the matched peptide) compared to non labeled, and this is greater for the lower collision energy. In all cases the proportion is even greater for decoy peptides.

Figure (16): Overview of amount of fragment ion signal associated with unassigned complementary b/y pairs. In terms of both number of fragments and proportion of total signal, the TMT label increases the amount of complementary b/y pairs. This phenomenon is more prevalent at the lower collision energy

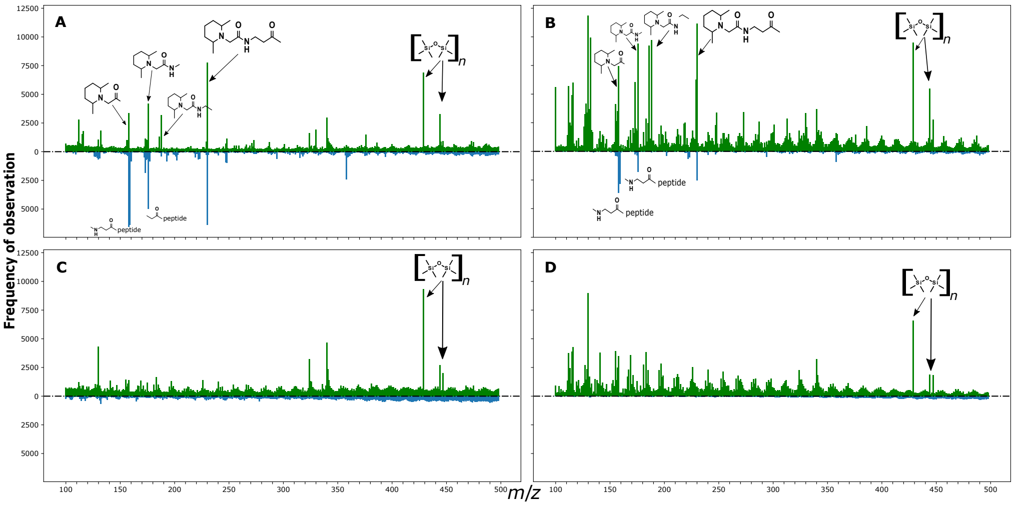


Figure (17): Histogram of counts of fragment ions observed in all four datasets. The *m/z* value of the fragment is given on the *x*-axis and the frequency on the *y*-axis. Shown in green are the counts for all fragment ions and in blue (with counts descending from the origin) are the complementary fragments (parent mass minus given *m/z* assuming a +2 parent charge state). The upper panel shows the TMT labeled peptides (A, 35% nCE; B, 45% nCE), the lower panel shows the unlabeled peptide fragments (C, 35% nCE; D, 45% nCE). The TMT labeled peptides yield a large proportion of fragment ions which are complementary to a set of low *m/z* fragments which correspond to unexpected cleavage of the TMT reporter group. Likely chemical structures are given for the complementary and low *m/z* peaks, originating from cleavage of the TMT-tag at unexpected positions. Low *m/z* ions with no complementary partners that are present in both labeled and unlabeled datasets correspond to polydimethylcyclosiloxane ions (nominal masses: 429 & 445; used for lock-mass determination), and immonium ions.

there is a much more pronounced difference between TMT and unlabeled spectra, as seen in figure 16a. For target spectra at both collision energies the mean number of unassigned pairs per spectrum is around 2.5 (slightly lower for 45% nCE), compared to less than 0.5 for unlabeled data. Overall, the decoy spectra have many more unassigned b/y pairs than the targets, but consistently show higher numbers of pairs coming from TMT labeled sets. Viewing these data as a proportion of TIC (Fig. 16b) we observe that the mean signal covered by unassigned pairs exhibits a four-fold increase between the labeled and unlabeled samples for target peptides. This difference is lower for decoy peptides. These observations suggest that the addition of the TMT tag leads to unexpected complementary ions occurring during peptide fragmentation.

The investigation now turned towards the unassigned (non-backbone) fragment ions. All peptide matches with an FDR below 1% in both TMT and unlabeled data sets were profiled - the unassigned fragment ions and also their charged-loss equivalents (calculated by subtracting the mass of each fragment at charge state +1 from an assumed parent mass of +2) were plotted for *m/z* 100 to 500 for each data set. Figure 17, with the low mass ions counting up from the origin and the charged loss *m/z*'s descending from the *x*-axis, shows that TMT labeled peptides yield many frequently observed complementary ions compared to very few in the non-labeled data set, in agreement with the increased complementary pairs seen in the previous section. Three to four groups of these complementary ions are present. The higher collision energy also produces a larger number of frequently occurring, unmatched, low

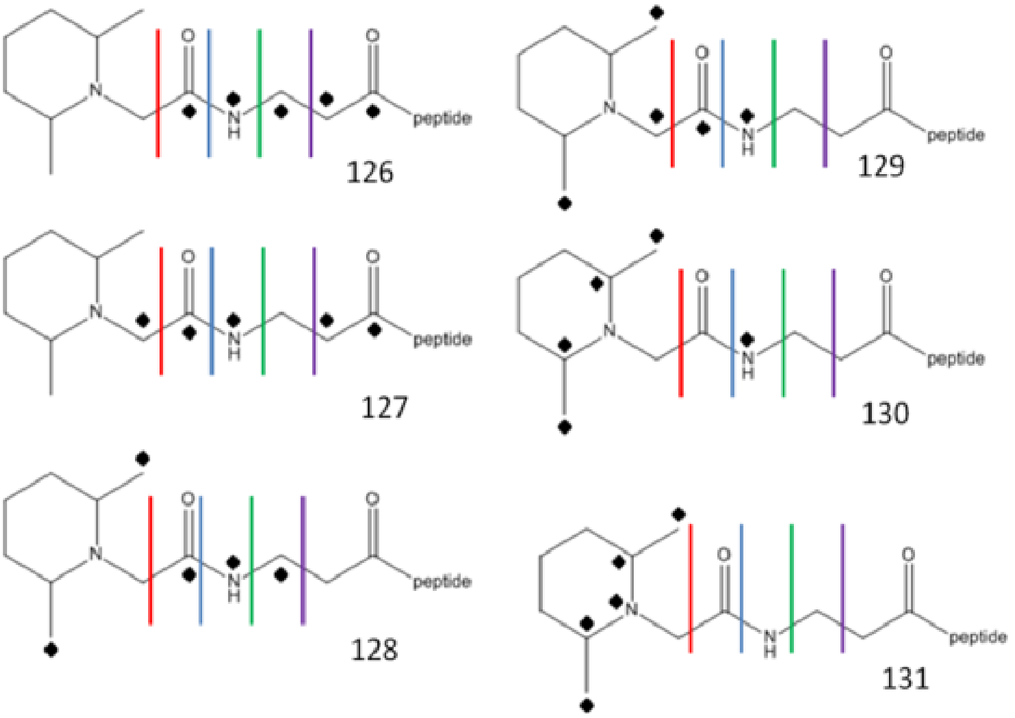


Figure (18): Depiction of possible sites of fragmentation within the TMT reporter group three of which are unexpected. The red line denotes the intended location of the HCD cleavage site. The blue line shows the location of fragmentation resulting in the addition of CO (the ’155 series’) to the reporter ions, green the addition of CNHO (the ’175 series’) and C_2NH_3O (the ’186 series’). For ease of understanding only the original TMT6-plex is shown. The black diamonds represent the heavy C or N atoms distributed across the moiety.

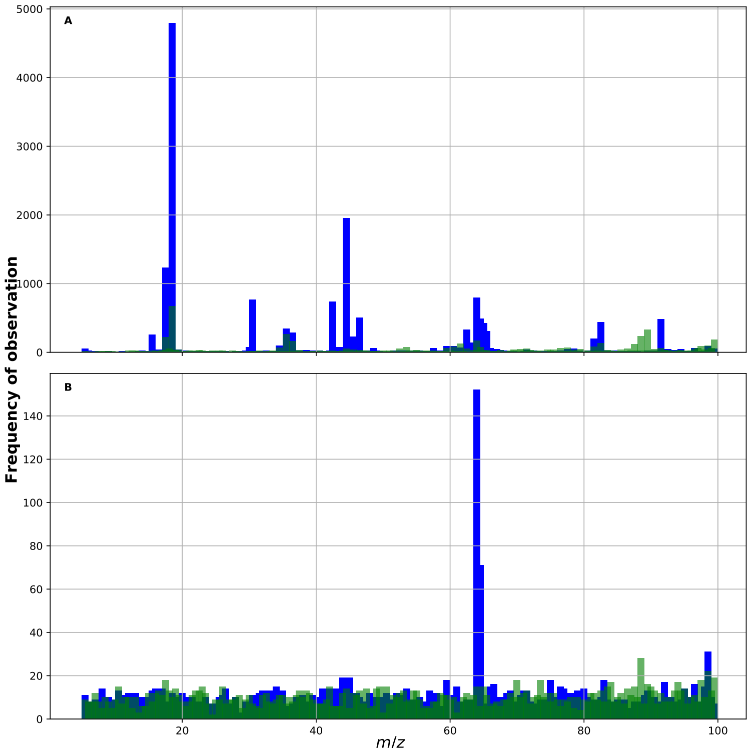
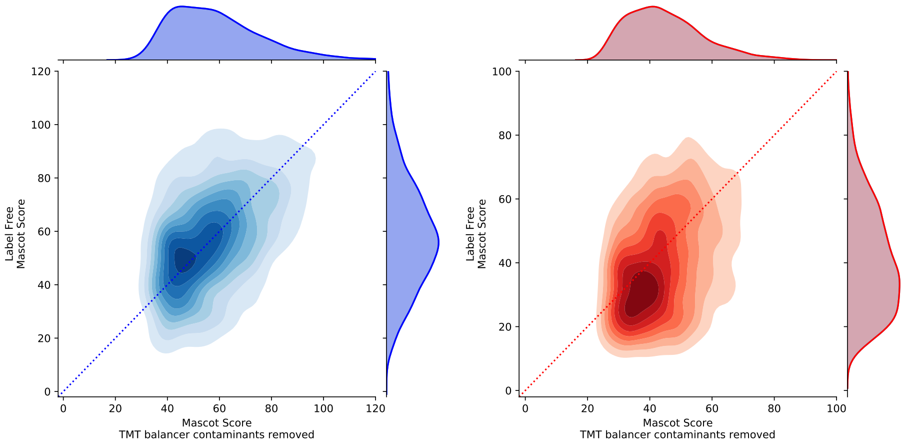


Figure (19): Zoom in on unassigned fragments (blue in Fig. 17) for the *m/z* region up to 100. Upper panel (A) is 35% nCE, lower (B) is 45% nCE. shown in blue are the TMT associated fragments in green are the unlabeled. TMT seems mass fragments than the lower energy with four or five maxima found in all datasets. These can be attributed to immonium ions, the lock mass peptide polydimethylcyclosiloxane (nominal mass: 429 & 445) and two unknown peaks with nominal masses around 320 & 340. Present in both the complementary and low mass plots we see the *m/z*'s corresponding to the TMT tag itself at 230.17 and the parent mass minus 230.17, a set of peaks corresponding to the reporter ion and the mass of carbon monoxide (as reported by Pichler *et al.*[249]), another set of peaks corresponding to a further cleavage along the TMT balancer group around *m/z* 175. A third set of peaks at approximately *m/z* 186 corresponds to a third break in the TMT balancer group and is present almost exclusively in the low mass ions. The deconvolution and deisotoping step of isobarQuant leads to the apparent disappearance of some of the reporter ion-associated charged loss fragments. In the 35% nCE TMT sample there are a few more maxima found: a series of peaks around *m/z* 202, 219 and 250. These correspond to the fragments from a neutral loss of 63 Da from oxidized methionine plus CO from the TMT balancer group, plus CHNO from the balancer group and lastly plus carbon monoxide from the balancer group with two oxidized methionines and two neutral losses. This finding could explain the low abundance of ions associated with methionine in the normalized intensity plots per cleavage site (Figs.13 & 14). Present only in the charged loss fragment ions is a peak corresponding to the loss of TMT labeled lysine from the parent ion (at both fragmentation energies) at around *m/z* 358. Focusing solely on the low *m/z* region of the complementary fragments (5-100 *m/z*) we observe a substantial increase in losses for TMT labeled peptides, mainly at 35% nCE. These cluster around regions corresponding to neutral losses such as the mass of water (-18), ammonia (-17) and from oxidized methionine (-64). We also find a peak at 45 *m/z* (corresponding to -COOH loss from the C-terminus of the peptide) and 30 *m/z* (corresponding to decarboxylation of aspartic acid or glutamic acid). There are also peaks of unknown origin present at *m/z* 43, 82 & 91 seen solely in the lowest collision energy, TMT-labeled data set.

The findings above further illustrate that the addition of a TMT reporter group affects the fragmentation of peptides, and this effect is more striking at the lower collision energy. More b-ions are detected because of the stabilizing effect of the TMT-group itself. This in turn leads to fewer internal fragments. One might expect that the increased presence of b-ions would translate into improved Mascot scores, but in fact the opposite is often the case. One reason for this might indeed be the high intensity peaks resulting from cleavage of the TMT label at unexpected points within the balancer group outlined above. To investigate this, all ions associated with this balancer cleavage were removed from the .mgf files created by isobarQuant and a further round of Mascot searches was performed. Following filtering at 1% peptide FDR, scores of shared TMT-labeled peptides increased to a level much more similar to their label-free counterparts (Fig. 20). For the 35% nCE collision energy there was a slight increase in mean overall score of 0.25 (Fig. 20a), and for the higher collision energy the effect of removing the TMT contaminant peaks resulted in an increase in Mascot ion score of 3.67 for TMT labeled peptides compared to the unlabeled ones



(a) 35% nCE

(b) 45% nCE

Figure (20): Two-dimensional density plots of Mascot score shifts on peptides shared between TMT labeled and and unlabeled (label free (LF)) data following removal of fragments coming from unexpected cleavage of the TMT-balancer group. On the left panel (blue) data for the lower collision energy (35% nCE) and on the right panel (red) data for the higher (45% nCE). On the *x*-axis the Mascot scores for peptides following TMT-contaminant fragment removal, on the *y*-axis the equivalent scores for the unlabeled peptides. Comparing this to untreated data (Fig. 7) we see scores nearly returned to the level of their unlabeled counterparts.

(Fig. 20b). It is worth noting at this point that the increase in Mascot score extended to all peptide matches, for both target and decoy (see table 7), actually leading to a derease in numbers of PSMs passing the 1% FDR filter.

The increase in Mascot scores following removal of TMT contaminant peaks piqued my interest to see if the stabilizing effect of the TMT tag on b-ions could in some way be harnessed to improve general scoring of PSMs and in particular to improve on the H-score algorithm published a few years before[165] and to try to increase the separation between target and decoy peptides. The first attempts were similar to those of Nielsen *et al.*[250] and tried to take advantage of the increased number of b-ions by awarding additional points to PSMs where the number of complementary b/y pairs was above a given threshold. This approach had no improvements over the standard H-score, because any b/y pairs found had essentially already been utilized in fragment cleavage sites included in the existing H-score. A second approach, this time with a focus on penalizing spectra with one or more unassigned b/y pairs based on the assumption that more unassigned b/y pairs are present in decoy peptides (Fig. 16a) was considered, but also rejected because the presence of one or more unassigned b/y pairs does not necessarily make any assertion about the correctness of the given spectrum to the peptide match (particularly when the maximum H-score has already been attained), but rather states the potential chimeracy of the MS/MS spectrum. Whilst this information cannot be employed to improve the H-score algorithm, it may be useful to give an indication of a peptide's co-elution status.

	1% FDR score thresh-old	PSMs < 1% FDR thresh-old	total count target (decoy)	mean score target (decoy)
35% standard processing	28	7175	17092 (4370)	26 (7)
35% TMT-balancer contaminants removed	35	6252 (-13%)	18112 (5190)	27 (9)
45% standard processing	24	9443	18343 (3138)	24 (7)
45% TMT-balancer contaminants removed	26	9067 (- 4%)	18863 (3473)	25 (7)

Table (7): The increase in Mascot score brought about by removal of fragments resulting from unexpected cleavage of TMT-balancer groups can lead to a small increase in the number of PSMs passing 1% FDR

3.2.2 Chimeric spectrum estimation using unassigned complementary b/y pairs

As described above in section 1.3, TMT-based quantification can suffer when more than one precursor (and consequently two or more sets of reporter ions) are produced and analyzed together. This has largely been solved by the S2I filtering and correction approach developed and included in isobarQuant (described above in section 2.3.1), however, in cases where several precursor masses are very close or identical within the given tolerance, additional insight about the presence of potential chimeric spectra could be gained by assessing the number of unassigned b/y pairs. The sixteen TMT-labeled files were interrogated and the precursor S2I value was plotted against the number of unassigned b/y pairs (any pairs originating from unexpected TMT-balancer group fragments were excluded for this analysis). There is no correlated relationship between the S2I value and the number of unassigned b/y pairs, rather the increased number of spectra with the given S2I value yields higher numbers of unassigned b/y pairs. The implication of this could be that a small number of seemingly clean spectra (with high S2I values and therefore low assumed ratio compression) still suffer from co-eluting peptides as evidenced by pairs of ions whose sum equals that of the precursor mass (within the given tolerance) but which are not present in the backbone fragments of the given peptide. Using this kind of approach to uncover chimeric spectra was also suggested by Gorshkov and co-workers[251]. Once again it was relatively easy to find examples of this using the data stored in the .hdf5 files. From this we see that one spectrum with an S2I of 1.0 was actually pervaded with co-eluted peaks as shown in figure 22. It is necessary to note that the values calculated here reflect the search tolerance and not the isolation window of the selected precursor used in the experiment. Applying this value would likely reveal more complementary fragments.

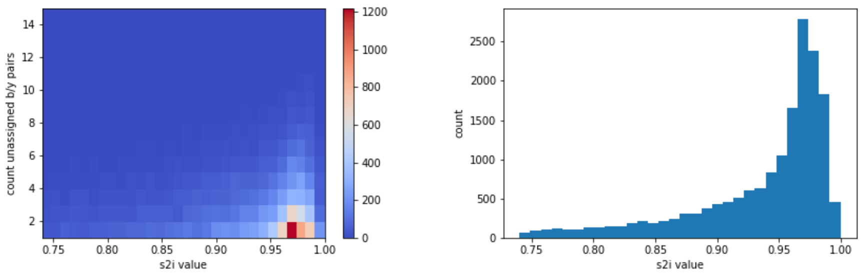


Figure (21): The number of unassigned b/y pairs plotted against the S2I value for all sixteen TMT labeled samples with at least one unassigned complementary pair. The number of unassigned b/y pairs is not correlated to the S2I value itself but rather to the total number of spectra at a given S2I value. The second plot (right) shows the distribution of all spectra with an S2I value greater than 0.75 and includes spectra with no unassigned b/y pairs.

3.2.3 H-score improvements using contiguous explained ratio

The attempts outlined above to try and improve H-score using the counts of b/y pairs failed to increase the separation between decoy and target peptides despite higher numbers of unassigned b/y pairs in decoy peptides. However, the increased number of b (or y) ions can still be useful in scoring; not as pairs but in the form of a ladder of contiguously explained cleavage sites. The original H-score took only the sum total of explained cleavage sites into account, irrespective of their location within the peptide. For a target peptide, where a fragment ion match is not solely due to chance, we would expect a greater proportion of ions to be located adjacent to one another. This contrasts with situation in decoy peptides where matches to a non-existent decoy peptide sequence can only occur by random chance and thus the likelihood of finding several explained fragment ion sites adjacent to one another is much reduced. By taking the maximum number of contiguous explained sites within the peptide and dividing this by the total number of possible sites (to correct for different length peptides) one can obtain a ratio which should increase the separation between targets and decoys better than using the total count alone. This calculated, 'contiguous' ratio for the decoy and target peptides acquired at the different collision energies is displayed in figure 23 and shows that, indeed, a higher proportion of decoy peptides have a contiguous explained ratio of less than 0.6, in contrast to the large number of target peptides with a ratio of 1.0. The overall increased number of both target and decoy peptides present with the TMT-labeled is again evident (see also Fig. 4). On the left side of each graph we observe that in TMT-datasets the contiguous ratio categories of 0 or 0.1 are made up almost entirely of decoy spectra. This fact was incorporated into an adapted H-score with the aim of achieving a better separation between decoy and target peptides than with the current implementation. The basis for the H-score should still be the total number of explained cleavage sites and the reward of three bonus points should still be given to peptides where all sites are explained (a contiguous ratio of 1.0) but the one-point bonus for peptides with all but one site explained will be replaced by a bonus of two points for all peptides with an explained contiguous ratio of greater than 0.1. Figure 24 shows the greater number of peptides below the 1% FDR threshold for theTMT-datasets (an addition of around

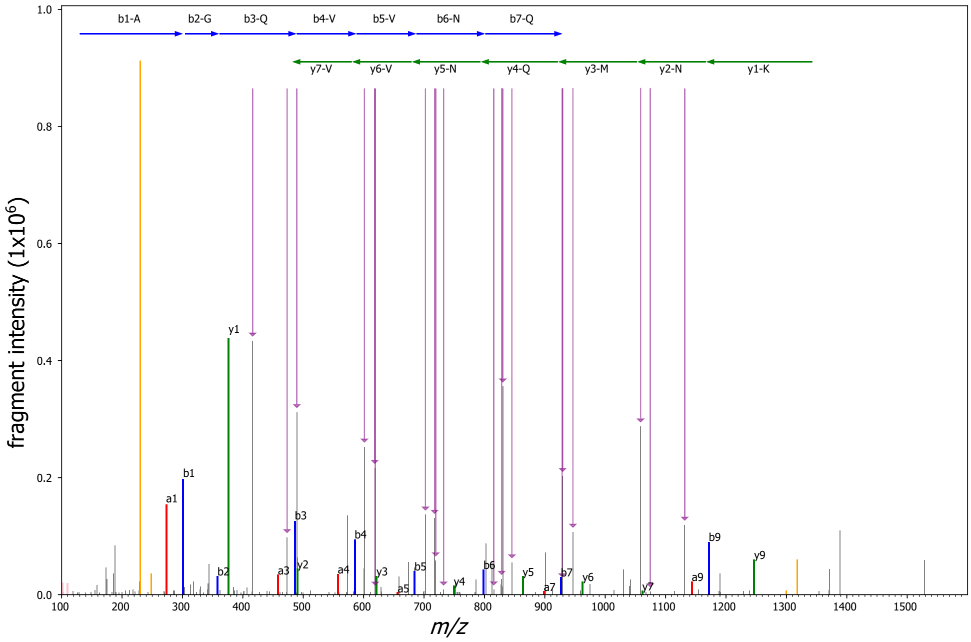


Figure (22): Graphical representation (generated directly from .hdf5 file) of a highly chimeric spectrum, matched by Mascot to sequence AGQVVNQMNK, score 16, with an S2I value of 1.0 (no apparent co-elution). Gray bars are acquired, deconvoluted and deisotoped fragment ions. B-ions are overlaid in blue, y-ions in green and a-ions in red. All TMT-derived fragments are shown in orange. The purple arrows show all 20 fragment peaks from possible chimeric spectra. The Mascot ions score is not particularly good (presumably due to the chimeracy of the spectrum) but could still be used for quantification. There is no indication from the S2I value that a co-elution event has occurred.

1,000 peptides over all eight TMT samples) with a much more modest improvement for non-labeled peptides.

4 Discussion

By interrogating the .hdf5 files created by isobarQuant (where the interpreted, searched data are stored alongside the raw acquired spectra) it has been possible to see that the addition of a TMT label to peptides affects the not only the size and number of peptides which are matched, but also the actual fragmentation of the peptides inside the mass spectrometer. The addition of a TMT label not only changes the fragmentation pattern by yielding reporter ions in the low *m/z* region of the spectrum (as per its primary intent) but also leads to the stabilization of and resulting increase in b-type fragment ions. This stabilization is also evidenced by a reduction in internal and immonium ions for labeled peptides and seems to result in an increase in numbers of target peptides identified, but also an increase in the number of decoy peptides and hence no significant change in FDR . On average the retention time of a TMT-labeled peptide is increased by between 10 and 12 minutes.

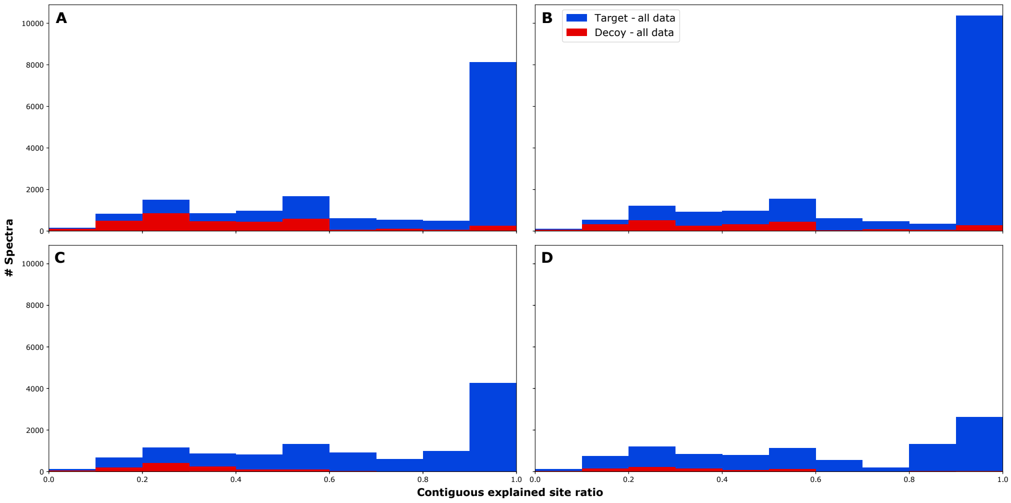


Figure (23): Histogram of all rank 1 target and decoy peptides binned by contiguous explained site ratio at all four collision energies with no peptide length cut off. The upper panel displays the values for TMT labeled peptides (35% nCE [A] and 45% nCE [B]) and the lower panel the values for the unlabeled peptides (35% nCE [C] and 45% nCE [D]). There are more target peptides with higher ratios of contiguous explained sites. The TMT-labeled samples have a larger number of spectra with fully contiguous explained sites. There are more decoy peptides with explained contiguous site ratios below 0.6 and a larger number overall with TMT-labeled peptides (compare with Fig. 4).The totals on the far left (ratio 0 to 0.1) are better represented by the decoy peptides

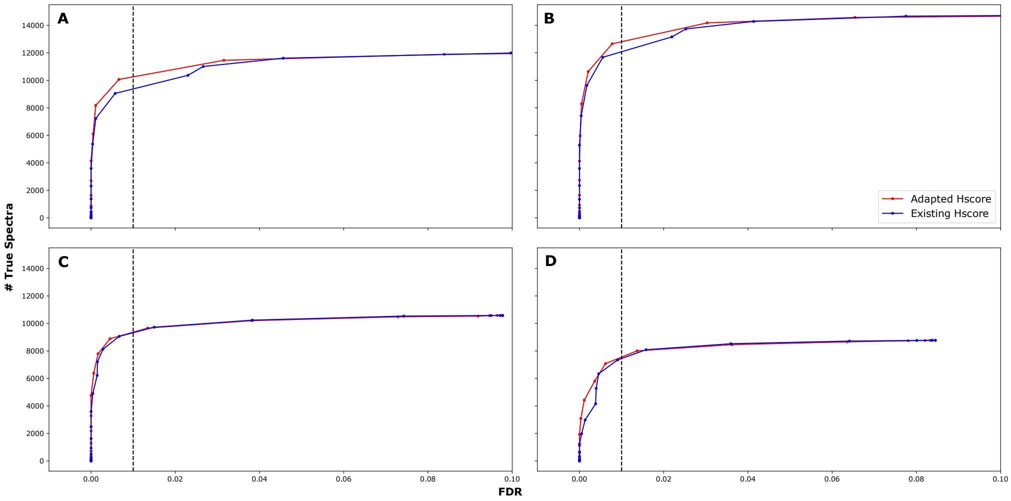


Figure (24): The H-score algorithm was adapted to take the contiguous explained ratio into account: rewarding two extra points if the ratio is greater than 0.1 rather than issuing an extra point if all but one of the cleavages sites were explained. The upper panel displays the values for TMT labeled peptides (35% nCE [A] and 45% nCE [B]) and the lower panel the values for the unlabeled peptides (35% nCE [C] and 45% nCE [D]). The black dashed line denotes the 1% FDR

4.1 Increased b-ion stability and small changes in cleavage bias

The increased presence of the b-ions and much lower a-ion intensities suggests that the labeled b-type fragments are more stable and less prone to degradation than their non-labeled counterparts. This is likely due to sequestration of one proton on the basic TMT-label at the N-terminus of the peptide. This stabilization prevents the mobile proton from being able to initiate further fragmentation events, which in turn leads to much less degradation to lower *b*- and *a*-ions[248] and also to fewer internal and immonium ions. This effect is less pronounced in samples acquired at a higher collision energy lending protons increased mobility. The increased b-ion stability leads to an overall increase in coverage by b-ion ladders but not too great an increase in overall coverage. It is tantalizing to state that the increased b-ion stability brought about by the addition of an inexpensive TMT-zero group (i.e. a non-isotopically labeled TMT group with no added heavy atoms) could be useful in studies to localize PTMs where additional b-ion ladder information can be used. It may also be useful in studies of the mobile proton itself. The benefits of the TMT label in phospho-proteomic studies has already been demonstrated by Jiang *et al.*[252].

The overall increased number of decoy peptide identifications compared to the increase in target peptides associated with TMT labeling was a source of quite some frustration during this project. A large number of TMT decoy peptides have TMT-reporter ions and thanks to the increased b-ion fragments a greater chance of matching to random PSMs in the search file.

4.2 Change in Mascot scores and unexpected TMT-balancer fragmentation

A swell in the number of b-ion fragments does translate into an increase in complementary b/y pairs, but this increase is by amplified by spurious b/y complementary pairs resulting from the unexpected cleavage of the TMT tag. These tag-related contaminant peaks are present at relatively high abundances and have a deleterious effect on the Mascot peptide scoring (the algorithm penalizes high-intensity signals which cannot be accounted for by any of the expected peptide fragments), leading to a general reduction in Mascot scores. After investigating the source and nature of these contaminant peaks and then removing them from the .mgf files prior to searching, we observe that the Mascot scores for TMT peptides also identified in unlabeled samples is returned to a level similar to that observed for the unlabeled counterparts. Overall, Mascot scores increase on average by between four and eight points following removal of TMT-contaminants, but this score increase is not accompanied by any significant change in FDR , again because the proportion of targets to decoy peptides remains more or less unchanged. This removal step would therefore only be necessary in labs where peptides are excluded based on empirically-determined Mascot score cut offs derived from label-free experiments, otherwise exclusion by an FDR threshold should suffice.

4.3 Identification of chimeric spectra and H-score improvements

The increased presence of complementary pairs in target peptides compared to decoys cannot be incorporated into the existing H-score. An additional reward for presence of complementary b/y pairs does not add any new information to that which is already contained in the H-score (number of explained cleavage sites). Despite the higher number of unassigned b/y pairs observed in decoy peptides, an attempt to penalize these fragments does not improve the separation between decoy and peptide matches. This is because the number of unassigned b/y pairs does not perform any assessment of the match between the suggested peptide and the identified fragments but rather evaluates the presence or absence of another, co-eluting or chimeric, spectrum. There is potential to use this value to exclude highly-chimeric peptides from being used for protein quantification in a way analogous to the S2I filtering, since the co-eluting reporter ions can lead to dampening of the true quantification signal, or ratio compression. This kind of approach has been proposed by Gorshkov *et al.*[251] who use it in the context of label-free quantification. However the calculation of complementary fragments should take the isolation window into account rather than the search tolerance used.

The use of the contiguous explained ratio in the H-score algorithm was shown to increase the number of PSMs passing the 1% FDR threshold. The random nature of fragment ion matches in decoy peptides means that explained sites adjacent to one another are less likely, which is also valid for any incorrect PSM assignment. The increase in b-ions with TMT-labeled peptides leads to an increased likelihood of longer contiguous explained ratios (from the N-terminus) and therefore an increase in the potential to perform better amino acid localization. This H-score improvement could be implemented straight away and will be more beneficial for TMT-labeled peptides than unlabeled but equally did not show any deterioration in performance for unlabeled peptides.

Currently the future of scoring algorithms or peptide fragmentation prediction seems to lie in the hands of AI and machine learning. While writing this thesis, several publications came out where the authors aimed to predict retention times or peptide fragmentation patterns using machine learning approaches on very large datasets of acquired and searched mass spectrometry data[108–110,253]. It will be interesting to see how these methods perform and evolve, and if isobarQuant can be used in conjunction with any of them; since the identification, quantification and raw data are all kept in a single file, indexed to allow fast access.

5 Author contribution to project

Toby Mathieson designed and performed the processing of all experiments in this section using isobarQuant; the laboratory experiments were performed and mass spectra were acquired at Cellzome GmbH, a GSK Company. The author carried out all steps of the data analysis including extracting relevant data and visualizations.

Part V

General discussion and outlook

1 isobarQuant: Discussion

This thesis presents and describes a stand-alone tool, isobarQuant, that is able to process, manipulate and quantify raw mass spectrometry data originating from any of the family of Thermo Fisher Scientific Orbitrap mass spectrometers. Its development and application to different problems has been thoroughly described and discussed in the three main parts of this thesis.

For a single .raw file its output comprises three text files (one protein-level, one peptide-level and one summary) and one .hdf5 file. The quantification of isobarically labeled or precursor labeled peptides is performed at the individual peptide level and peptides and quantification values are then filtered, corrected and extrapolated to the protein level. The .hdf5 file allows raw data to be stored alongside the interpretations made using it. This output is easy to access by many different programming languages. The text outputs follow a standardized format irrespective of the method of quantification used. It is written in Python and would therefore be platform-independent were it not for the requirement of the vendor software to run under the Windows operating system. This means that, at least, the pre-Mascot part of the pipeline is tied to that platform, although several possible workarounds have recently opened and would need to be further explored.

isobarQuant was the first software tool to provide a method for the correction of isobaric-label quantification values from the potential influence of co-eluting peptides (ratio compression) and is, to the best of the author's knowledge, the only tool to employ a bootstrapping method to determine a level of confidence in the protein fold changes when based on more than a given number of (isobarically labeled) PSMs (default value: four). It was shown to outperform MaxQuant in the determination of accurate peptide fold changes when a high level of background interference (from co-eluting peptides) is present and introduced a second filter criterion P2T to exclude the reporter ions of peptides whose signals are very close to the instrument noise level. Other software tools calculate a value similar to the S2I, (the PIF of MaxQuant, for example), but isobarQuant is unique in that it interpolates between two MS/MS scans to derive the value at the precursor selection, not just in the preceding one, a method that was later published in the realm of metabolites in 2017[254]. A relatively recent investigation into the effects of ratio compression using a set of ground-truth phospho peptides[255] showed that the S2I is not the only metric that one can use to measure potential co-elution of peptides and that even peptides with high S2I values still suffered from some ratio compression. They also stated that the Andromeda score was (loosely) inversely proportional to the level of S2I, which makes sense and could apply to the Mascot search engine as well, since any evidence of co-eluting peptides will be observable in the fragment ions of the MS/MS spectrum and thereby lower the score. This has been reflected in isobarQuant's use of additional (peptide) filter criteria (such as Mascot score or delta_to_next) since its inception.

isobarQuant is able to perform multiple rounds of peptide- and protein-level quantification for optimization / investigation of different parameters (e.g. using different filters and cutoffs) without having to re-perform the first (time consuming) raw data extraction and processing steps. Each new round of quantification creates a new, distinctly-named set of outputs. This was critical in the development of the MS1-based quantification of SILAC labeled peptides in protein half-life determination where three additional filter criteria were investigated and optimized to exclude peptides with inaccurate fold changes. The first, the 'prior-ion ratio' was a novel metric introduced to filter out peptide (pairs) with an unexpectedly high-intensity peak present in light or heavy peptides at the *m/z* corresponding to the loss of one neutron from the monoisotopic ion. The presence of such a peak at this position indicates that a co-eluting (interfering) isotope cluster is present and should not be used in downstream analyses. The second optimization was in the filtering procedure to gain the highest number of quantified peptides with the greatest accuracy based on the least-squares fit of the exact model to the observed isotopic distribution. This was the first software to use an exact model, rather than averagine to determine SILAC peptide fold changes. Thirdly, a minimum peptide count threshold used to exclude indeterminable peptide ratios from the calculation of protein quantification was established. These additions to the software led to the publication of several SILAC-pulse labeled datasets cataloging the protein half-lives of four human and one murine primary cell lines. It has been demonstrated in this report that isobarQuant is able to give a higher number of accurate, very low peptide fold changes when compared to MaxQuant and the resulting protein half-lives have given some valuable insights into the differences in turnover of protein complexes and their components across these five different cell lines. The use of this resource as a source of non-perturbed protein half-lives will be of great value in many different types of study; ranging from the effects of compounds on protein turnover to the assessment of the efficacy (at the protein level) of CRISPR gene knockdowns.

The .hdf5 output of isobarQuant was used directly to interrogate and analyze spectra coming from TMT labeled reagents. The changes which the large protophilic group brought about in terms of fragmentation were readily extracted and visualized using a Python API. This quick and easy data access allowed for the investigation into the altered fragmentation properties of TMT-labeled peptides, an extension to the work done by Pichler *et al.* regarding unexpected TMT balancer group fragmentation and the effect on Mascot scoring[249]. It also allowed an improvement in the existing H-score algorithm as well as providing a potential method to flag TMT labeled peptides with a higher-than-estimated amount of co-elution (and ratio compression).

During the creation of this report, where the results of isobarQuant were compared to those of MaxQuant a small weakness with the MaxQuant software was uncovered: namely that quantification was not possible with some more recently acquired .raw files, possibly those with a slightly different acquisition method. These did not pose a problem for isobarQuant and quantification was possible without any intervention, but the fact that the MaxQuant software is closed-source prevented the author from being able to effectively troubleshoot the issue or establish where the problem with the software (or raw file) lay. Attempts were made with multiple versions of the software, including the most recent. The issue was luckily solvable by using files acquired on an earlier version of the Thermo Fisher Scientific Xcalibur software. isobarQuant's code is fully open-source and were similar problems with processing files to arise, it would be simple to pinpoint where exactly the issue was.

isobarQuant was the basis for several TMT-labeled, TPP experiments for elucidation of potential off-targets and has provided a straight-forward way to process TMT-labeled data including S2I / ratio compression correction.

2 isobarQuant software: Outlook

As described in chapter II, the software would benefit from some refactoring to enable it to be more modular. This might also make its functionality more amenable to other node-based workflow tools. It would be very desirable to support the results of more than one search engine (Mascot) and to allow a more Ursgal[127] / Percolator[124]-like approach in order to derive the most from the results of different search engines and the acquired data at the highest accuracy. The direct incorporation of H-score, likely with the improvements described above (4.3), could be envisioned in the very near future.

A general concern for species that are not well annotated, for which the uniqueness calculation is hampered by many overlapping proteins (which lack sufficient annotation stating that proteins are translated from the same gene and are consequently not grouped and quantified together), it might be beneficial to adapt isobarQuant to use a solution similar to that provided by Ursgal and assign the name of the peptide group as a concatenation of the names of the constituent peptides. Groups of peptides with the concatenated name could then be quantified together. This will enable different proteoforms to be separately quantified and could be extended to allow quantification of individual peptides or peptide groups in a way similar to that described by Zecha *et al.*[256], which is essential if we consider that different proteoforms of the same gene may be differently expressed and turned-over in different tissues or under different conditions and that indeed all multi-exon genes have been shown to undergo alternative splicing[257].

There are currently two potential ways to remove isobarQuant's dependency on the Window operating system. The first, as mentioned above is to use the freely-available Windows emulator software layer for Linux, WINE. One would be then be able to use isobarQuant on either Linux or Windows. There has also been a method published (https://pypi.org/project/pythonnet) which provides a Python wrapper to the .NET Common Language Runtime that could interface directly with Thermo Fisher Scientific's Xcalibur library. It has already been used to visualize a chromatogram from a Thermo Fisher Scientific instrument and could therefore represent an alternative way to provide this.

One feature that isobarQuant is not currently benefiting from is the possibility to perform MS1 and MS/MS re-calibration at the *m/z* and RT level. This capability has actually been programmed, but has not yet been deployed because of issues associated with connectivity to and interaction with the Mascot server / search engine and the configuration of individual searches. This would need to be overcome if this should take place without a good deal of manual user interaction. Whilst the lack of re-calibration for poorly calibrated runs will affect the numbers of peptide identifications made within a narrow precursor search tolerance, the tolerance can be increased without too high an increase in false positives and since isobarQuant uses and a best-cohort selection method to determine the reporter ions based on supplied quantification masses, even a poorly-calibrated instrument run can still deliver accurate quantification. As long as an adequate FDR cutoff is used, the data need not be discarded.

2.1 Further improvements to S2I and isotope impurity corrections

A couple of recent reports have suggested new ways to calculate a more accurate S2I value. Iwasaki and co-workers suggest extending the scope of the S2I calculation to include points farther along the XIC than just the preceding and succeeding MS/MS event[207], which could potentially mitigate some of the issues raised by by Hogrebe *et al.* where even MS precursors with apparently high S2I values still yield co-eluting reporter ions in their fragment spectra[255]. Searle and Yergey have published a new method to carry out isotopic correction, using a linear algebra approach that is common in electrical engineering and an improvement to the S2I correction implemented in isobarQuant (ref.[171]) by preventing over correction of interference when high reporter signal stems from a single channel[208].

2.2 Support for new instrumentation

In mid-2019 Thermo Fischer Scientific released another instrument in the Orbitrap family. The Orbitrap Exploris 480 boasts higher scan speeds and increased resolution, such as the addition of the ΦSDM[258] capability. Because this instrument has a slightly different API, a small adaptation to isobarQuant will be required in order to process files acquired form it. However, no fundamental alterations to how the software operates will be needed and code changes should take no longer than one or two days to implement. Another innovation recently applied in the arena of shotgun proteomics is the addition of the high-field asymmetricwaveform ion mobility spectrometry (FAIMS) interface[259], which employs a form of ion mobility separation using alternative low and high electric fields before ions enter the orifice of the instrument. Ions of different mobilities are transmitted in turn by scanning the compensation voltage. This allows the removal of singly charged ions, interfering isobaric precursor species and works in concert with any type of MS/MS acquisition. FAIMS has been shown to improve quantification[260,261] and to allow identification with much shorter gradients[262]. Support for these kinds of experiments with isobarQuant might require more investment in time, since the nature of the acquisition is different, scanning through the alternative compensation voltages and creating a different experiment for each (each of which would require separate XIC , S2I and P2T extraction methods).

2.3 isobarQuant and cohort analyses

The field of proteomics has transitioned over the last decade from being a largely qualitative, with read-outs consisting of a list of protein identifications to being quantitative within a single or short series of experiments where up to eleven different conditions are compared accurately and precisely. The challenge now facing modern proteomics is to shift this ability to quantify relative protein abundance across much larger cohorts, for example over hundreds or even thousands of samples in a pharmaceutical or clinical research setting. Label-free methods, where unlimited numbers of runs can be combined offer some hope of a solution, but these are typically dogged by reproducibility issues, in part because of missing internal standards to correct for quantitative variations arising from sample preparation and analytical process. Such technical variations can lead to inaccurate measurements, especially for low-abundance proteins, as well as high false-positives in discovering proteins with altered states[263]. Labeling methods offer some hope in this regard and with the extension of a TMT-like reagent to a 16plex capability (possibly multiplied by an MS1 label such as SILAC to further extend to 32plex) but having the possibility to measure more than this order of magnitude in a single experiment is limited (Neucode[203] encoding is also limited to 32plex). All these approaches will suffer from the stochastic nature of DDA in precursor selection in the different runs which can, in turn lead to under-sampling of low-abundance proteins[264]. This is compounded by the instruments' dynamic exclusion settings, which mean that the same MS/MS spectrum may not be selected in consecutive runs, also reducing reproducibility. The DIA approach is attempting to overcome these issues and recently, a number of new pipelines were developed to support these efforts. PECAN[265], DIA-Umpire[266], and DirectDIA in Spectronaut™, Pulsar[267]. isobarQuant is not designed for that type of work flow, but it can certainly help to provide a consensus spectral library with ease as demonstrated in Part IV of this report.

The approach offering the most accurate and precise solution might actually lie in being able to combine the results of isobarically labeled experiments. This is itself not without problems, as highlighted recently by Brenes *et al.*[268]when analyzing and combining 24 TMT-10plex, MS3-quantified samples. The authors state that integrating two or more runs already leads to an increase in missing values (from <2% to ~25%), which goes against one of the primary benefits of using a TMT labeling strategy. They also note that using an MS2-based, TMT quantification method should reduce the batch effect, although in a DDA approach there is no guarantee that missing values will not be encountered. They go on to underline the importance of including a common control sample within each TMT batch against which to normalize across all samples and reduce batch effects. Of course, if the proteins of critical interest to the study are known *a priori*, a sensible approach might be to construct a targeted data acquisition (TDA) list of precursor *m/z's* , charge states and RTs for the instrument to fragment and quantify over all runs, based on one or two path finding experiments. isobarQuant could play a role in the creation of these lists, enabling the selection of the best set of peptides for the given protein(s) in terms of ease of identification (or so-called fly-ability) and also intensity of reporter ions generated. This could also be extended to select the most intense fragment ions from a peptide for use in a PRM (or SRM) approach.

Exactly how isobarQuant develops into the future remains to be seen. The different direct contributions that it has made so far to the Proteomics Community have been outlined in the three main parts within this report.

References

1. Wasinger, V. C. *et al.* Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *Electrophoresis* **16,** 1090–1094. doi:10.1002/elps.11501601185 (1995).
2. Cox, J. & Mann, M. Is Proteomics the New Genomics? *Cell* **130,** 395–398. doi:10.1016/j.cell.2007.07.032 (2007).
3. Thomson, J. J. Bakerian Lecture: Rays of Positive Electricity. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **89,** 1–20. doi:10.1098/rspa.1913.0057 (1913).
4. Aston, F. LXXIV. A positive ray spectrograph. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **38,** 707–714. doi:10.1080/14786441208636004 (2009).
5. McLafferty, F. W. Mass Spectrometric Analysis: Molecular Rearrangements. *Analytical Chemistry* **31,** 82–87. doi:10.1021/ac60145a015 (1959).
6. Biemann, K., Seibl, J. & Gapp, F. Mass spectrometric identification of amino acids. *Biochemical and Biophysical Research Communications* **1,** 307–311. doi:10.1016/0006-291X(59)90044-0 (1959).
7. Macek, B., Waanders, L. F., Olsen, J. V. & Mann, M. Top-down protein sequencing and MS3 on a hybrid linear quadrupole ion trap-orbitrap mass spectrometer. *Molecular and Cellular Proteomics* **5,** 949–958. doi:10.1074/mcp.T500042-MCP200 (2006).
8. Auclair, J. R. *et al.* Artifacts to avoid while taking advantage of top-down mass spectrometry based detection of protein S-thiolation. *Proteomics* **14,** 1152–1157. doi:10.1002/pmic.201300450 (2014).
9. Auclair, J. R. *et al.* Post-translational modification by cysteine protects Cu/Zn-superoxide dismutase from oxidative damage. *Biochemistry* **52,** 6137–6144. doi:10.1021/bi4006122 (2013).
10. Ayaz-Guner, S. *et al.* In vivo phosphorylation site mapping in mouse cardiac troponin I by high resolution top-down electron capture dissociation mass spectrometry: Ser22/23 are the only sites basally phosphorylated. *Biochemistry* **48,** 8161–8170. doi:10.1021/bi900739f (2009).
11. Skinner, O. S. *et al.* Top-down characterization of endogenous protein complexes with native proteomics. *Nature Chemical Biology* **14,** 36–41. doi:10.1038/nchembio.2515 (2018).
12. Olsen, J. V., Ong, S. E. & Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Molecular and Cellular Proteomics* **3,** 608–614. doi:10.1074/mcp.T400003-MCP200 (2004).

13. Quan, Q. *et al.* Fully Automated Multidimensional Reversed-Phase Liquid Chromatography with Tandem Anion/Cation Exchange Columns for Simultaneous Global Endogenous Tyrosine Nitration Detection, Integral Membrane Protein Characterization, and Quantitative Proteomics Mappin. *Analytical Chemistry* **87,** 10015–10024. doi:10.1021/acs.analchem.5b02619 (2015).

14. Bjellqvist, B. *et al.* Isoelectric focusing in immobilized pH gradients: Principle, methodology and some applications. *Journal of Biochemical and Biophysical Methods* **6,** 317–339. doi:10.1016/0165-022X(82)90013-6 (1982).

15. Alpert, A. J. Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. *Journal of Chromatography A* **499,** 177–196. doi:10.1016/S0021-9673(00)96972-3 (1990).

16. Zhang, X., Ye, J., Jensen, O. N. & Roepstorff, P. Highly efficient phosphopeptide enrichment by calcium phosphate precipitation combined with subsequent IMAC enrichment. *Molecular and Cellular Proteomics* **6,** 2032–2042. doi:10.1074/mcp.M700278-MCP200 (2007).

17. Possemato, A. P. *et al.* Multiplexed Phosphoproteomic Profiling Using Titanium Dioxide and Immunoaffinity Enrichments Reveals Complementary Phosphorylation Events. *Journal of Proteome Research* **16,** 1506–1514. doi:10.1021/acs.jproteome.6b00905 (2017).

18. Choudhary, C. *et al.* The growing landscape of lysine acetylation links metabolism and cell signalling. *Nature Reviews Molecular Cell Biology* **15,** 536–550. doi:10.1038/nrm3841 (2014).

19. Udeshi, N. D., Mertins, P., Svinkina, T. & Carr, S. A. Large-scale identification of ubiquitination sites by mass spectrometry. *Nature Protocols* **8,** 1950–1960. doi:10.1038/nprot.2013.120 (2013).

20. Glish, G. L. & Vachet, R. W. The basics of mass spectrometry in the twenty-first century. *Nature Reviews Drug Discovery* **2,** 140–150. doi:10.1038/nrd1011 (2003).

21. Fenn, J. B. *et al.* Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246,** 64–71. doi:10.1126/science.2675315 (1989).

22. Hillenkamp, F., Karas, M., Beavis, R. C. & Chait, B. T. Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry of Biopolymers. *Analytical Chemistry* **63,** 1193A–1203A. doi:10.1021/ac00024a716 (1991).

23. So, P. K., Hu, B. & Yao, Z. P. Mass spectrometry: Towards in vivo analysis of biological systems. *Molecular BioSystems* **9,** 915–929. doi:10.1039/c2mb25428j (2013).

24. Gross, J. H. *Mass spectrometry: A textbook: Second edition* 1–753. doi:10.1007/978-3-642-10711-5 (2011).

25. Chernushevich, I. V., Loboda, A. V. & Thomson, B. A. An introduction to quadrupole-time-of-flight mass spectrometry. *Journal of Mass Spectrometry* **36,** 849–865. doi:10.1002/jms.207 (2001).

26. Garabedian, A. *et al.* Towards Discovery and Targeted Peptide Biomarker Detection Using nanoESI-TIMS-TOF MS. *Journal of the American Society for Mass Spectrometry* **29,** 817–826. doi:10.1007/s13361-017-1787-8 (2018).

27. Makarov, A. Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry* **72,** 1156–1162. doi:10.1021/ac991131p (2000).

28. Makarov, A., Denisov, E., Lange, O. & Horning, S. Dynamic Range of Mass Accuracy in LTQ Orbitrap Hybrid Mass Spectrometer. *Journal of the American Society for Mass Spectrometry* **17,** 977–982. doi:10.1016/j.jasms.2006.03.006 (2006).

29. Olsen, J. V. *et al.* Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a C-trap. *Molecular and Cellular Proteomics* **4,** 2010–2021. doi:10.1074/mcp.T500030-MCP200 (2005).

30. Kingdon, K. H. A method for the neutralization of electron space charge by positive ionization at very low gas pressures. *Physical Review* **21,** 408–418. doi:10.1103/PhysRev.21.408 (1923).

31. Knight, R. D. Storage of ions from laser-produced plasmas. *Applied Physics Letters* **38,** 221–223. doi:10.1063/1.92315 (1981).

32. Michalski, A. *et al.* Mass spectrometry-based proteomics using Q exactive, a high-performance benchtop quadrupole orbitrap mass spectrometer. *Molecular and Cellular Proteomics* **10,** M111.011015. doi:10.1074/mcp.M111.011015 (2011).

33. Olsen, J. V. *et al.* A dual pressure linear ion trap orbitrap instrument with very high sequencing speed. *Molecular and Cellular Proteomics* **8,** 2759–2769. doi:10.1074/mcp.M900375-MCP200 (2009).

34. ThermoFisher Scientific. *Thermo Fisher :: Planet Orbitrap* 2019.

35. McAlister, G. C. *et al.* MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Analytical Chemistry* **86,** 7150–7158. doi:10.1021/ac502040v (2014).

36. Erickson, B. K. *et al.* Evaluating multiplexed quantitative phosphopeptide analysis on a hybrid quadrupole mass filter/linear ion trap/orbitrap mass spectrometer. *Analytical Chemistry* **87,** 1241–1249. doi:10.1021/ac503934f (2015).

37. Erickson, B. K. *et al.* A Strategy to Combine Sample Multiplexing with Targeted Proteomics Assays for High-Throughput Protein Signature Characterization. *Molecular Cell* **65,** 361–370. doi:10.1016/j.molcel.2016.12.005 (2017).

38. Roepstorff, P. & Fohlman, J. Letter to the editors - Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biological Mass Spectrometry* **11,** 601. doi:10.1002/bms.1200111109 (1984).

39. Mitchell Wells, J. & McLuckey, S. A. *Collision-induced dissociation (CID) of peptides and proteins* in *Methods in Enzymology* 148–185 (2005). doi:10.1016/S0076-6879(05)02005-7.

40. Schwartz, J. C. *High-Q pulsed fragmentation in ion traps* 2006.

41. Wu, W. W. *et al.* Identification of proteins and phosphoproteins using pulsed Q collision induced dissociation (PQD). *Journal of the American Society for Mass Spectrometry* **22,** 1753–1762. doi:10.1007/s13361-011-0197-6 (2011).

42. Olsen, J. V. *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods* **4,** 709–712. doi:10.1038/nmeth1060 (2007).

43. Murray, K. K. *et al.* Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013). *Pure and Applied Chemistry* **85,** 1515–1609. doi:10.1351/PAC-REC-06-04-06 (2013).

44. Michalski, A., Neuhauser, N., Cox, J. & Mann, M. A systematic investigation into the nature of tryptic HCD spectra. *Journal of Proteome Research* **11,** 5479–5491. doi:10.1021/pr3007045 (2012).

45. Zubarev, R. A. Electron-capture dissociation tandem mass spectrometry. *Current Opinion in Biotechnology* **15,** 12–16. doi:10.1016/j.copbio.2003.12.002 (2004).

46. Syka, J. E. P. *et al.* Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 9528–9533. doi:10.1073/pnas.0402700101 (2004).

47. Murray, K. *File:Peptide fragmentation.gif - Wikimedia Commons*

48. Bruderer, R. *et al.* Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Molecular and Cellular Proteomics* **16,** 2296–2309. doi:10.1074/mcp.RA117.000314 (2017).

49. Sinitcyn, P., Rudolph, J. D. & Cox, J. Computational Methods for Understanding Mass Spectrometry-Based Shotgun Proteomics Data. *Annual Review of Biomedical Data Science* **1,** 207–234. doi:10.1146/annurev-biodatasci-080917-013516 (2018).

50. Silva, J. C. *et al.* Absolute Quantification of Proteins by LCMS E. *Molecular & Cellular Proteomics* **5,** 144–156. doi:10.1074/mcp.m500230-mcp200 (2006).

51. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–342. doi:10.1038/nature10098 (2011).

52. Wisniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Molecular and Cellular Proteomics* **13,** 3497–3506. doi:10.1074/mcp.M113.037309 (2014).

53. Lundgren, D. H., Hwang, S. I., Wu, L. & Han, D. K. Role of spectral counting in quantitative proteomics. *Expert Review of Proteomics* **7,** 39–53. doi:10.1586/epr.09.69 (2010).

54. Pavelka, N. *et al. Statistical similarities between transcriptomics and quantitative shotgun proteomics data* in *Molecular and Cellular Proteomics* **7** (2008), 631–644. doi:10.1074/mcp.M700240-MCP200.

55. Cox, J. *et al.* Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Molecular & Cellular Proteomics* **13,** 2513–2526. doi:10.1074/MCP.M113.031591 (2014).

56. Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology* **17,** 994–999. doi:10.1038/13690 (1999).

57. Yao, X. *et al.* Proteolytic 18O labeling for comparative proteomics: Model studies with two serotypes of adenovirus. *Analytical Chemistry* **73,** 2836–2842. doi:10.1021/ac001404c (2001).

58. Hsu, J. L., Huang, S. Y., Chow, N. H. & Chen, S. H. Stable-Isotope Dimethyl Labeling for Quantitative Proteomics. *Analytical Chemistry* **75,** 6843–6852. doi:10.1021/ac0348625 (2003).

59. Hsu, J. L., Huang, S. Y. & Chen, S. H. Dimethyl multiplexed labeling combined with microcolumn separation and MS analysis for time course study in proteomics. *Electrophoresis* **27,** 3652–3660. doi:10.1002/elps.200600147 (2006).

60. Kristensen, D. B. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular and Cellular Proteomics* **1,** 376–386 (2002).

61. Geiger, T. *et al.* Super-SILAC mix for quantitative proteomics of human tumor tissue Supplementary figures and text : Supplementary Figure 1. *Nature Methods* **7,** 383–385. doi:10.1038/nmeth.1446 (2010).

62. Hebert, A. S. *et al.* Neutron-encoded mass signatures for multiplexed proteome quantification. *Nature Methods* **10,** 332–334. doi:10.1038/nmeth.2378 (2013).

63. Sleno, L. The use of mass defect in modern mass spectrometry Special Feature : Tutorial The use of mass defect in modern mass spectrometry. *Journal of mass spectrometry* **47,** 226–236. doi:10.1002/jms.2953 (2015).

64. Ross, P. L. *et al.* Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular and Cellular Proteomics* **3,** 1154–1169. doi:10.1074/mcp.M400129-MCP200 (2004).

65. Thompson, A. *et al.* Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry* **75,** 1895–1904. doi:10.1021/ac0262560 (2003).

66. Werner, T. *et al.* High-resolution enabled TMT 8-plexing. *Analytical chemistry* **84,** 7188–7194. doi:10.1021/ac301553x (2012).

67. Bantscheff, M. *et al.* Robust and sensitive iTRAQ quantification on an LTQ Orbitrap mass spectrometer. *Molecular & cellular proteomics : MCP* **7,** 1702–13. doi:10.1074/mcp.M800029-MCP200 (2008).

68. Saw, Y. O. *et al.* iTRAQ underestimation in simple and complex mixtures: "The good, the bad and the ugly". *Journal of Proteome Research* **8,** 5347–5355. doi:10.1021/pr900634c (2009).

69. Winter, S. V. *et al.* EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification. *Nature Methods* **15,** 527–530. doi:10.1038/s41592-018-0037-8 (2018).

70. Mann, M., Hojrup, P. & Roepstorff, P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol Mass Spectrom* **22,** 338–345. doi:10.1002/bms.1200220605 (1993).

71. DJ, P., P, H. & AJ, B. Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology* **3,** 327–332 (1993).

72. Henzel, W. J. *et al. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases* in *Proceedings of the National Academy of Sciences of the United States of America* **90** (National Academy of Sciences, Department of Protein Chemistry, Genentech, Inc., South San Francisco, CA 94080-4990., 1993), 5011–5015. doi:10.1073/pnas.90.11.5011.

73. Eidhammer, I., Flikka, K., Martens, L. & Mikalsen, S. O. *Computational Methods for Mass Spectrometry Proteomics* 1–284. doi:10.1002/9780470724309 (John Wiley & Sons, 2007).

74. Tabb, D. L., Saraf, A. & Yates, J. R. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Analytical chemistry* **75,** 6415–6421 (2003).

75. Kapp, E. A. *et al.* Mining a Tandem Mass Spectrometry Database to Determine the Trends and Global Factors Influencing Peptide Fragmentation. *Analytical Chemistry* **75,** 6251–6264. doi:10.1021/ac034616t (2003).

76. Savitski, M. M. *et al.* Evaluation of data analysis strategies for improved mass spectrometry-based phosphoproteomics. *Analytical Chemistry* **82,** 9843–9849. doi:10.1021/ac102083q (2010).

77. Wenger, C. D. & Coon, J. J. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. *Journal of Proteome Research* **12,** 1377–1386. doi:10.1021/pr301024c (2013).

78. Risk, B. A., Edwards, N. J. & Giddings, M. C. A peptide-spectrum scoring system based on ion alignment, intensity, and pair probabilities. *Journal of Proteome Research* **12,** 4240–4247. doi:10.1021/pr400286p (2013).

79. Fenyo, D. & Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical chemistry* **75,** 768–774 (2003).

80. Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research* **10,** 1794–1805. doi:10.1021/pr101065j (2011).

81. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **5,** 976–989. doi:10.1016/1044-0305(94)80016-2 (1994).

82. Searle, B. C. *et al.* Identification of protein modifications using MS/MS de novo sequencing and the OpenSea alignment algorithm. *Journal of Proteome Research* **4,** 546–554. doi:10.1021/pr049781j (2005).

83. Bafna, V. & Edwards, N. SCOPE: A probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* **17,** S13–21. doi:10.1093/bioinformatics/17.suppl_1.S13 (2001).

84. Perkins, D. N., Pappin, D. J. C., Creasy, D. M. & Cottrell, J. S. Probability based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS: An International Journal* **20,** 3551–3567 (1999).

85. Nesvizhskii, A. I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods* **4,** 787–797. doi:10.1038/nmeth1088 (2007).

86. Vaudel, M. *et al.* D-score: A search engine independent MD-score. *Proteomics* **13,** 1036–1041. doi:10.1002/pmic.201200408 (2013).

87. Savitski, M. M. *et al.* Confident phosphorylation site localization using the mascot delta score. *Molecular and Cellular Proteomics* **10,** M110. 003830. doi:10.1074/mcp.M110.003830 (2011).

88. Beausoleil, S. A. *et al.* A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature Biotechnology* **24,** 1285–1292. doi:10.1038/nbt1240 (2006).

89. Chick, J. M. *et al.* A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat Biotechnol* **33,** 743–749. doi:10.1038/nbt.3267 (2015).

90. Savitski, M. M., Nielsen, M. L. & Zubarev, R. A. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Molecular and Cellular Proteomics* **5,** 935–948. doi:10.1074/mcp.T500034-MCP200 (2006).

91. Creasy, D. M. & Cottrell, J. S. Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics* **2,** 1426–1434. doi:10.1002/1615-9861(200210)2:10<1426::AID-PROT1426>3.0.CO;2-5 (2002).

92. Kong, A. T. *et al.* MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods* **14,** 513–520. doi:10.1038/nmeth.4256 (2017).

93. Zolg, D. P. *et al.* Building ProteomeTools based on a complete synthetic human proteome. *Nature Methods* **14,** 259–262. doi:10.1038/nmeth.4153 (2017).

94. Craig, R., Cortens, J. C., Fenyo, D. & Beavis, R. C. Using annotated peptide mass spectrum libraries for protein identification. *Journal of Proteome Research* **5,** 1843–1849. doi:10.1021/pr0602085 (2006).

95. Frewen, B. E. *et al.* Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Analytical Chemistry* **78,** 5678–5684. doi:10.1021/ac060279n (2006).

96. Deutsch, E. W., Lam, H. & Aebersold, R. PeptideAtlas: A resource for target selection for emerging targeted proteomics workflows. *EMBO Reports* **9,** 429–434. doi:10.1038/embor.2008.56 (2008).

97. Vizcaíno, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research* **44,** D447–D456. doi:10.1093/nar/gkv1145 (2016).

98. Martens, L. *et al.* PRIDE: The proteomics identifications database. *Proteomics* **5,** 3537–3545. doi:10.1002/pmic.200401303 (2005).

99. Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7,** 655–667. doi:10.1002/pmic.200600625 (2007).

100. Ma, B. *et al.* PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry* **17,** 2337–2342. doi:10.1002/rcm.1196 (2003).

101. Frank, A. & Pevzner, P. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling - Analytical Chemistry (ACS Publications). *Anal Chem* **77,** 964–973 (2005).

102. Johnson, R. S. & Taylor, J. A. Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Applied Biochemistry and Biotechnology - Part B Molecular Biotechnology* **22,** 301–315. doi:10.1385/MB:22:3:301 (2002).

103. Chi, H. *et al.* pNovo: de novo peptide sequencing and identification using HCD spectra. *Journal of proteome research* **9,** 2713–2724 (2010).

104. Jeong, K., Kim, S. & Pevzner, P. A. UniNovo: A universal tool for de novo peptide sequencing. *Bioinformatics* **29,** 1953–1962. doi:10.1093/bioinformatics/btt338 (2013).

105. Guthals, A., Clauser, K. R., Frank, A. M. & Bandeira, N. Sequencing-grade de novo analysis of MS/MS triplets (CID/HCD/ETD) from overlapping peptides. *Journal of Proteome Research* **12,** 2846–2857. doi:10.1021/pr400173d (2013).

106. Degroeve, S., Martens, L. & Jurisica, I. MS2PIP: A tool for MS/MS peak intensity prediction. *Bioinformatics* **29,** 3199–3203. doi:10.1093/bioinformatics/btt544 (2013).

107. Degroeve, S., Maddelein, D. & Martens, L. MS2PIP prediction server: Compute and visualize MS2 peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Research* **43,** W326–W330. doi:10.1093/nar/gkv542 (2015).

108. C Silva, A. S., Bouwmeester, R., Martens, L. & Degroeve, S. Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *Bioinformatics* (ed Wren, J.) doi:10.1093/bioinformatics/btz383 (2019).

109. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods* **16,** 509–518. doi:10.1038/s41592-019-0426-7 (2019).

110. Tiwary, S. *et al.* High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature Methods* **16,** 519–525. doi:10.1038/s41592-019-0427-6 (2019).

111. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57,** 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x (1995).

112. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4,** 207–214. doi:10.1038/nmeth1019 (2007).

113. Bianco, L., Mead, J. A. & Bessant, C. Comparison of novel decoy database designs for optimizing protein identification searches using ABRF sPRG2006 standard MS/MS data sets. *Journal of Proteome Research* **8,** 1782–1791. doi:10.1021/pr800792z (2009).

114. Wang, G. *et al.* Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. *Analytical Chemistry* **81,** 146–159. doi:10.1021/ac801664q (2009).

115. Käll, L., Storey, J. D., MacCoss, M. J. & Noble, W. S. Posterior error probabilities and false discovery rates: Two sides of the same coin. *Journal of Proteome Research* **7,** 40–44. doi:10.1021/pr700739d (2008).

116. Choi, H. & Nesvizhskii, A. I. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *Journal of Proteome Research* **7,** 47–50. doi:10.1021/pr700747q (2008).

117. Navarro, P. & Vazquez, J. A refined method to calculate false discovery rates for peptide identification using decoy databases. *Journal of Proteome Research* **8,** 1792–1796. doi:10.1021/pr800362h (2009).

118. Keich, U., Tamura, K. & Noble, W. S. Averaging Strategy To Reduce Variability in Target-Decoy Estimates of False Discovery Rate. *Journal of Proteome Research* **18,** 585–593. doi:10.1021/acs.jproteome.8b00802 (2019).

119. Lam, H., Deutsch, E. W. & Aebersold, R. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. *Journal of Proteome Research* **9,** 605–610. doi:10.1021/pr900947u (2010).

120. Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics* **73,** 2092–2123. doi:10.1016/j.jprot.2010.08.009 (2010).

121. Vaudel, M. *et al.* PeptideShaker enables reanalysis of MS-derived proteomics data sets: To the editor. *Nature Biotechnology* **33,** 22–24. doi:10.1038/nbt.3109 (2015).

122. Kwon, T. *et al.* MSblender: A probabilistic approach for integrating peptide identifications from multiple database search engines. *Journal of Proteome Research* **10,** 2949–2958. doi:10.1021/pr2002116 (2011).

123. Nahnsen, S. *et al.* Probabilistic consensus scoring improves tandem mass spectrometry peptide identification. *Journal of Proteome Research* **10,** 3332–3343. doi:10.1021/pr2002879 (2011).

124. Käll, L. *et al.* Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods* **4,** 923–925. doi:10.1038/nmeth1113 (2007).

125. Wen, B. *et al.* IPeak: An open source tool to combine results from multiple MS/MS search engines. *Proteomics* **15,** 2916–2920. doi:10.1002/pmic.201400208 (2015).

126. Vaudel, M. *et al.* SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **11,** 996–999. doi:10.1002/pmic.201000595 (2011).

127. Kremer, L. P. *et al.* Ursgal, Universal Python Module Combining Common Bottom-Up Proteomics Tools for Large-Scale Analysis. *Journal of Proteome Research* **15,** 788–794. doi:10.1021/acs.jproteome.5b00860 (2016).

128. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry abilities that proteins are present in a sample on the basis. *Analytical chemistry* **75,** 4646–4658. doi:10.1021/ac0341261 (2003).

129. Reiter, L. *et al.* Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Molecular and Cellular Proteomics* **8,** 2405–2417. doi:10.1074/mcp.M900317-MCP200 (2009).

130. Savitski, M. M. *et al.* A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Molecular and Cellular Proteomics* **14,** 2394–2404. doi:10.1074/mcp.M114.046995 (2015).

131. Serang, O. & Noble, W. S. Faster mass spectrometry-based protein inference: junction trees are more efficient than sampling and marginalization by enumeration. *IEEE/ACM Trans Comput Biol Bioinform* **9,** 809–817. doi:10.1109/TCBB.2012.26 (2012).

132. Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: The protein inference problem. *Molecular and Cellular Proteomics* **4,** 1419–1440. doi:10.1074/mcp.R500012-MCP200 (2005).

133. Huang, T., Wang, J., Yu, W. & He, Z. Protein inference: A review. *Briefings in Bioinformatics* **13,** 586–614. doi:10.1093/bib/bbs004 (2012).

134. The, M. *et al.* A Protein Standard That Emulates Homology for the Characterization of Protein Inference Algorithms. *Journal of Proteome Research* **17,** 1879–1886. doi:10.1021/acs.jproteome.7b00899 (2018).

135. Li, Y. F. & Radivojac, P. Computational approaches to protein inference in shotgun proteomics. *BMC bioinformatics* **13 Suppl 1,** S4. doi:10.1186/1471-2105-13-S16-S4 (2012).

136. Li, Y. F. *et al.* A bayesian approach to protein inference problem in shotgun proteomics. *Journal of Computational Biology* **16,** 1183–1193. doi:10.1089/cmb.2009.0018 (2009).

137. Tabb, D. L. *et al.* Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *Journal of Proteome Research* **9,** 761–776. doi:10.1021/pr9006365 (2010).

138. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26,** 1367–1372. doi:10.1038/nbt.1511 (2008).

139. Sinitcyn, P. *et al.* MaxQuant goes Linux. *Nature Methods* **15,** 401. doi:10.1038/s41592-018-0018-y (2018).

140. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* **13,** 731–740. doi:10.1038/nmeth.3901 (2016).

141. Deutsch, E. W. *et al.* A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **10,** 1150–1159. doi:10.1002/pmic.200900375 (2010).

142. Pfeuffer, J. *et al.* OpenMS - A platform for reproducible analysis of mass spectrometry data. *Journal of Biotechnology* **261,** 142–148. doi:10.1016/j.jbiotec.2017.05.016 (2017).

143. KNIME.com GmbH. *About KNIME* 2018.

144. ThermoFisher Scientific. *Proteome Discoverer Software* 2017.

145. MacLean, B. *et al.* Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26,** 966–968. doi:10.1093/bioinformatics/btq054 (2010).

146. Blank, C. *et al.* Disseminating metaproteomic informatics capabilities and knowledge using the galaxy-P framework. *Proteomes* **6,** 7. doi:10.3390/proteomes6010007 (2018).

147. Vizcaíno, J. A. *et al.* The Proteomics Identifications (PRIDE) database and associated tools: Status in 2013. *Nucleic Acids Research* **41.** doi:10.1093/nar/gks1262 (2013).

148. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–587. doi:10.1038/nature13319. arXiv: 209 (2014).

149. Schmidt, T. *et al.* ProteomicsDB. *Nucleic acids research* **46,** D1271–D1281. doi:10.1093/nar/gkx1029 (2018).

150. Vizcaíno, J. A. *et al.* *ProteomeXchange provides globally coordinated proteomics data submission and dissemination* 2014. doi:10.1038/nbt.2839.

151. Deutsch, E. W. *et al.* The ProteomeXchange consortium in 2017: Supporting the cultural change in proteomics public data deposition. *Nucleic Acids Research* **45,** D1100–D1106. doi:10.1093/nar/gkw936 (2017).

152. Farrah, T. *et al.* PASSEL: The PeptideAtlas SRMexperiment library. *Proteomics* **12,** 1170–1175. doi:10.1002/pmic.201100515 (2012).

153. Okuda, S. *et al.* JPOSTrepo: An international standard data repository for proteomes. *Nucleic Acids Research* **45,** D1107–D1111. doi:10.1093/nar/gkw1080 (2017).

154. Ma, J. *et al.* Iprox: An integrated proteome resource. *Nucleic Acids Research* **47,** D1211–D1217. doi:10.1093/nar/gky869 (2019).

155. Savitski, M. M. *et al.* Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* **346,** 1255784. doi:10.1126/science.1255784 (2014).

156. Bantscheff, M. *et al.* Chemoproteomics profiling of HDAC inhibitors reveals selective targeting of HDAC complexes. *Nature Biotechnology* **29,** 255–268. doi:10.1038/nbt.1759 (2011).

157. Salvisberg, C. *et al.* Exploring the human tear fluid: Discovery of new biomarkers in multiple sclerosis. *Proteomics - Clinical Applications* **8,** 185–194. doi:10.1002/prca.201300053 (2014).

158. Tsuchida, S. *et al.* Application of quantitative proteomic analysis using tandem mass tags for discovery and identification of novel biomarkers in periodontal disease. *Proteomics* **13,** 2339–2350. doi:10.1002/pmic.201200510 (2013).

159. Andaya, A. *et al.* Phosphorylation stoichiometries of human Eukaryotic initiation factors. *International Journal of Molecular Sciences* **15,** 11523–11538. doi:10.3390/ijms150711523 (2014).

160. Jia, W., Andaya, A. & Leary, J. A. Novel mass spectrometric method for phosphorylation quantification using cerium oxide nanoparticles and tandem mass tags. *Analytical Chemistry* **84,** 2466–2473. doi:10.1021/ac203248s (2012).

161. Franken, H. *et al.* Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. *Nature Protocols* **10,** 1567–1593. doi:10.1038/nprot.2015.101 (2015).

162. HDF Group. *Introduction to HDF5* 2010.

163. Alted, F. *et al. PyTables User's Guide - PyTables 3.5.1 documentation* 2018.

164. Mujezinovic, N. *et al.* Cleaning of raw peptide MS/MS spectra: Improved protein identification following deconvolution of multiply charged peaks, isotope clusters, and removal of background noise. *PROTEOMICS* **6,** 5117–5131. doi:10.1002/pmic.200500928 (2006).

165. Savitski, M. M., Mathieson, T., Becher, I. & Bantscheff, M. H-score, a mass accuracy driven rescoring approach for improved peptide identification in modification rich samples. *Journal of Proteome Research* **9,** 5511–5516. doi:10.1021/pr1006813 (2010).

166. Reiz, B., Kertész-Farkas, A., Pongor, S. & Myers, M. P. Chemical rule-based filtering of MS/MS spectra. *Bioinformatics* **29,** 925–932. doi:10.1093/bioinformatics/btt061 (2013).

167. Köcher, T. *et al.* Altered Mascot search results by changing the m/z range of MS/MS spectra: Analysis and potential applications. *Analytical and Bioanalytical Chemistry* **400,** 2339–2347. doi:10.1007/s00216-010-4572-0 (2011).

168. Liu, H., Sadygov, R. G. & Yates, J. R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical Chemistry* **76,** 4193–4201. doi:10.1021/ac0498563 (2004).

169. Ow, S. Y. *et al.* Minimising iTRAQ ratio compression through understanding LC-MS elution dependence and high-resolution HILIC fractionation. *Proteomics* **11,** 2341–2346. doi:10.1002/pmic.201000752 (2011).

170. Werner, T. *et al.* Ion coalescence of neutron encoded TMT 10-plex reporter ions. *Analytical Chemistry* **86,** 3594–3601. doi:10.1021/ac500140s (2014).

171. Savitski, M. M. *et al.* Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *Journal of Proteome Research* **12,** 3586–3598. doi:10.1021/pr400098r (2013).

172. Savitski, M. M. *et al.* Targeted Data Acquisition for Improved Reproducibility and Robustness of Proteomic Mass Spectrometry Assays. *Journal of the American Society for Mass Spectrometry* **21,** 1668–1679. doi:10.1016/j.jasms.2010.01.012 (2010).

173. Savitski, M. M. *et al.* Delayed fragmentation and optimized isolation width settings for improvement of protein identification and accuracy of isobaric mass tag quantification on orbitrap-type mass spectrometers. *Analytical Chemistry* **83,** 8959–8967. doi:10.1021/ac201760x (2011).

174. Pachl, F., Fellenberg, K., Wagner, C. & Kuster, B. Ultra-high intra-spectrum mass accuracy enables unambiguous identification of fragment reporter ions in isobaric multiplexed quantitative proteomics. *Proteomics* **12,** 1328–1332. doi:10.1002/pmic.201100622 (2012).

175. Elias, J. E. & Gygi, S. P. *Target-decoy search strategy for mass spectrometry-based proteomics.* in *Methods in molecular biology (Clifton, N.J.)* 55–71 (2010). doi:10.1007/978-1-60761-444-9_5.

176. Hughes, C. S. *et al.* Ultrasensitive proteome analysis using paramagnetic bead technology. *Molecular Systems Biology* **10,** 757. doi:10.15252/msb.20145625 (2014).

177. Keshishian, H. *et al.* Multiplexed, quantitative workflow for sensitive biomarker discovery in plasma yields novel candidates for early myocardial injury. *Molecular and Cellular Proteomics* **14,** 2375–2393. doi:10.1074/mcp.M114.046813 (2015).

178. Kalxdorf, M., Eberl, H. C. & Bantscheff, M. *Monitoring dynamic changes of the cell surface glycoproteome by quantitative proteomics* in *Methods in Molecular Biology* 47–59 (Humana Press, New York, NY, 2017). doi:10.1007/978-1-4939-7201-2_3.

179. Mateus, A., Määttä, T. A. & Savitski, M. M. Thermal proteome profiling: unbiased assessment of protein state through heat-induced stability changes. *Proteome Science* **15,** 13. doi:10.1186/s12953-017-0122-4 (2016).

180. Mateus, A. *et al.* Thermal proteome profiling in bacteria: probing protein state inÂ vivo. *Molecular systems biology* **14,** e8242. doi:10.15252/msb.20188242 (2018).

181. Mathieson, T. *et al.* Systematic analysis of protein turnover in primary cells. *Nature Communications* **9.** doi:10.1038/s41467-018-03106-1 (2018).

182. Jethwa, A. *et al.* TRRAP is essential for regulating the accumulation of mutant and wild-type p53 in lymphoma. *Blood* **131,** 2789–2802. doi:10.1182/blood-2017-09-806679 (2018).

183. Panov, A. & Gygi, S. P. Analysis of Independent Differences (AID) detects complex thermal proteome profiles independent of shape and identifies candidate panobinostat targets. *bioRxiv,* 751818. doi:10.1101/751818 (2019).

184. Becher, I. *et al.* Pervasive Protein Thermal Stability Variation during the Cell Cycle. *Cell* **173,** 1495–1507.e18. doi:10.1016/j.cell.2018.03.053 (2018).

185. Colombo, M., Pessey, O. & Marcia, M. Topology and enzymatic properties of a canonical Polycomb repressive complex 1 isoform. *FEBS Letters* **593,** 1837–1848. doi:10.1002/1873-3468.13442 (2019).

186. Brancini, G. T. P., Ferreira, M. E. S., Rangel, D. E. N. & Braga, G. Ú. L. Combining Transcriptomics and Proteomics Reveals Potential Post-transcriptional Control of Gene Expression After Light Exposure in Metarhizium acridum. *G3&#58; Genes|Genomes|Genetics.* doi:10.1534/g3.119.400430 (2019).

187. Boos, F. *et al.* Mitochondrial protein-induced stress triggers a global adaptive transcriptional programme. *Nature cell biology* **21,** 442–451. doi:10.1038/s41556-019-0294-5 (2019).

188. Hao, Y. *et al.* Targetome analysis of chaperone-mediated autophagy in cancer cells. *Autophagy,* 1–14. doi:10.1080/15548627.2019.1586255 (2019).

189. Dai, L. *et al.* Horizontal Cell Biology: Monitoring Global Changes of Protein Interaction States with the Proteome-Wide Cellular Thermal Shift Assay (CETSA). *Annual Review of Biochemistry* **88,** 383–408. doi:10.1146/annurev-biochem-062917-012837 (2019).

190. Savitski, M. M. *et al.* Multiplexed Proteome Dynamics Profiling Reveals Mechanisms Controlling Protein Homeostasis. *Cell* **173,** 260–274.e25. doi:10.1016/j.cell.2018.02.030 (2018).

191. Perez-Perri, J. I. *et al.* Discovery of RNA-binding proteins and characterization of their dynamic responses by enhanced RNA interactome capture. *Nature Communications* **9.** doi:10.1038/s41467-018-06557-8 (2018).

192. Schulze, W. M. *et al.* Structural analysis of human ARS2 as a platform for cotranscriptional RNA sorting. *Nature Communications* **9.** doi:10.1038/s41467-018-04142-7 (2018).

193. Hassler, M. *et al.* Structural Basis of an Asymmetric Condensin ATPase Cycle. *Molecular Cell* **74,** 1175–1188.e9. doi:10.1016/j.molcel.2019.03.037 (2019).

194. Behrendt, A. *et al.* Asparagine endopeptidase cleaves tau at N167 after uptake into microglia. *Neurobiology of Disease* **130.** doi:10.1016/j.nbd.2019.104518 (2019).

195. Girstmair, H. *et al.* The Hsp90 isoforms from S. cerevisiae differ in structure, function and client range. *Nature Communications* **10.** doi:10.1038/s41467-019-11518-w (2019).

196. Odabasi, E. *et al.* Differential requirement for centriolar satellites in cilium formation among different vertebrate cells. *bioRxiv,* 478974. doi:10.1101/478974 (2018).

197. Odabasi, E., Gul, S., Kavakli, I. H. & Firat-Karalar, E. N. Centriolar satellites are required for efficient ciliogenesis and ciliary content regulation. *EMBO reports* **20.** doi:10.15252/embr.201947723 (2019).

198. Banzhaf, M. *et al.* The outer membrane lipoprotein NlpI nucleates hydrolases within peptidoglycan multi-enzyme complexes in Escherichia coli. *bioRxiv,* 609503. doi:10.1101/609503 (2019).

199. Wyllie, S. *et al.* Cyclin-dependent kinase 12 is a drug target for visceral leishmaniasis. *Nature* **560,** 192–197. doi:10.1038/s41586-018-0356-z (2018).

200. Sridharan, S. *et al.* Proteome-wide solubility and thermal stability profiling reveals distinct regulatory roles for ATP. *Nature Communications* **10.** doi:10.1038/s41467-019-09107-y (2019).

201. Obrdlik, A. *et al.* The Transcriptome-wide Landscape and Modalities of EJC Binding in Adult Drosophila. *Cell Reports* **28,** 1219–1236.e11. doi:10.1016/j.celrep.2019.06.088 (2019).

202. Mergner, J. *et al.* Mass-spectrometry-based draft of the Arabidopsis proteome. *Nature* **579,** 1–6. doi:10.1038/s41586-020-2094-2 (2020).

203. Rose, C. M. *et al.* Neutron encoded labeling for peptide identification. *Analytical Chemistry* **85,** 5129–5137. doi:10.1021/ac400476w (2013).

204. Merrill, A. E. *et al.* NeuCode labels for relative protein quantification. *Molecular and Cellular Proteomics* **13,** 2503–2512. doi:10.1074/mcp.M114.040287 (2014).

205. Braun, C. R. *et al.* Generation of Multiple Reporter Ions from a Single Isobaric Reagent Increases Multiplexing Capacity for Quantitative Proteomics. *Analytical Chemistry* **87,** 9855–9863. doi:10.1021/acs.analchem.5b02307 (2015).

206. Sonnett, M., Yeung, E. & Wuhr, M. Accurate, Sensitive, and Precise Multiplexed Proteomics Using the Complement Reporter Ion Cluster. *Analytical Chemistry* **90,** 5032–5039. doi:10.1021/acs.analchem.7b04713 (2018).

207. Iwasaki, M. *et al.* Removal of Interference MS/MS Spectra for Accurate Quantification in Isobaric Tag-Based Proteomics. *Journal of Proteome Research* **18,** 2535–2544. doi:10.1021/acs.jproteome.9b00078 (2019).

208. Searle, B. C. & Yergey, A. L. An efficient solution for resolving iTRAQ and TMT channel cross-talk. *Journal of Mass Spectrometry.* doi:10.1002/jms.4354 (2019).

209. Leufken, J. *et al.* PyQms enables universal and accurate quantification of mass spectrometry data. *Molecular and Cellular Proteomics* **16,** 1736–1745. doi:10.1074/mcp.M117.068007 (2017).

210. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications* **5,** 5277. doi:10.1038/ncomms6277 (2014).

211. Python Software Foundation. *PEP 373 – Python 2.7 Release Schedule | Python.org* 2008.

212. Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11,** 395. doi:10.1186/1471-2105-11-395 (2010).

213. Deutsch, E. mzML: A single, unifying data format for mass spectrometer output. *Proteomics* **8,** 2776–2777. doi:10.1002/pmic.200890049 (2008).

214. Hulstaert, N. *et al.* ThermoRawFileParser: modular, scalable and cross-platform RAW file conversion. *bioRxiv,* 622852. doi:10.1101/622852 (2019).

215. Adusumilli, R. & Mallick, P. *Data conversion with proteoWizard msConvert* in *Methods in Molecular Biology* 339–368 (2017). doi:10.1007/978-1-4939-6747-6_23.

216. Sturm, M. *et al.* OpenMS - An open-source software framework for mass spectrometry. *BMC Bioinformatics* **9,** 163. doi:10.1186/1471-2105-9-163 (2008).

217. Chambers, M. C. *et al.* An accessible proteogenomics informatics resource for cancer researchers. *Cancer Research* **77,** e43–e46. doi:10.1158/0008-5472.CAN-17-0331 (2017).

218. Meier, F. *et al.* BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nature Methods* **15,** 440–448. doi:10.1038/s41592-018-0003-5 (2018).

219. Goeminne, L. J., Argentini, A., Martens, L. & Clement, L. Summarization vs peptide-based models in label-free quantitative proteomics: Performance, pitfalls, and data analysis guidelines. *Journal of Proteome Research* **14,** 2457–2465. doi:10.1021/pr501223t (2015).

220. Savas, J. N. *et al.* Extremely long-lived nuclear pore proteins in the rat brain. *Science* **335,** 942. doi:10.1126/science.1217421 (2012).

221. Mitchell, C. J., Kim, M. S., Na, C. H. & Pandey, A. PyQuant: A versatile framework for analysis of quantitative mass spectrometry data. *Molecular and Cellular Proteomics* **15,** 2829–2838. doi:10.1074/mcp.O115.056879 (2016).

222. Senko, M. W., Beu, S. C. & McLafferty, F. W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry* **6,** 229–233. doi:10.1016/1044-0305(95)00017-8 (1995).

223. Becher, I. *et al.* Chemoproteomics reveals time-dependent binding of histone deacetylase inhibitors to endogenous repressor complexes. *ACS Chemical Biology* **9,** 1736–1746. doi:10.1021/cb500235n (2014).

224. Bantscheff, M. *et al.* Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nature Biotechnology* **25,** 1035–1044. doi:10.1038/nbt1328 (2007).

225. Kruse, U. *et al.* Chemoproteomics-based kinome profiling and target deconvolution of clinical multi-kinase inhibitors in primary chronic lymphocytic leukemia cells. *Leukemia* **25,** 89–100. doi:10.1038/leu.2010.233 (2011).

226. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols* **11,** 2301–2319. doi:10.1038/nprot.2016.136 (2016).

227. Drew, K. *et al.* Integration of over 9,000 mass spectrometry experiments builds a global map of human proteinÂ complexes. *Molecular Systems Biology.* doi:10.15252/msb.20167490 (2017).

228. Kang, U. B., Yeom, J., Kim, H. & Lee, C. Quantitative analysis of mTRAQ-labeled proteome using full MS scans. *Journal of Proteome Research* **9,** 3750–3758. doi:10.1021/pr9011014 (2010).

229. Wu, Y. *et al.* Five-plex isotope dimethyl labeling for quantitative proteomics. *Chemical Communications* **50,** 1708–1710. doi:10.1039/c3cc47998f (2014).

230. Christoforou, A. L. & Lilley, K. S. Isobaric tagging approaches in quantitative proteomics: The ups and downs. *Analytical and Bioanalytical Chemistry* **404,** 1029–1037. doi:10.1007/s00216-012-6012-9 (2012).

231. Lam, Y. W., Lamond, A. I., Mann, M. & Andersen, J. S. Analysis of Nucleolar Protein Dynamics Reveals the Nuclear Degradation of Ribosomal Proteins. *Current Biology* **17,** 749–760. doi:10.1016/j.cub.2007.03.064 (2007).

232. Piha, R. S., Cuénod, M. & Waelsch, H. Metabolism of histones of brain and liver. *Journal of Biological Chemistry* **241,** 2397–2404 (1966).

233. Toyama, B. H. *et al.* Identification of long-lived proteins reveals exceptional stability of essential cellular structures. *Cell* **154,** 971–982. doi:10.1016/j.cell.2013.07.037 (2013).

234. Rabut, G., Doye, V. & Ellenberg, J. Mapping the dynamic organization of the nuclear pore complex inside single living cells. *Nature Cell Biology* **6,** 1114–1121. doi:10.1038/ncb1184 (2004).

235. Kosinski, J. *et al.* Molecular architecture of the inner ring scaffold of the human nuclear pore complex. *Science* **352,** 363–365. doi:10.1126/science.aaf0643 (2016).

236. Lin, D. H. *et al.* Architecture of the symmetric core of the nuclear pore. *Science* **352.** doi:10.1126/science.aaf1015 (2016).

237. Price, J. C. *et al.* Analysis of proteome dynamics in the mouse brain. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 14508–14513. doi:10.1073/pnas.1006551107 (2010).

238. Daigle, N. *et al.* Nuclear pore complexes form immobile networks and have a very low turnover in live mammalian cells. *Journal of Cell Biology* **154,** 71–84. doi:10.1083/jcb.200101089 (2001).

239. Webster, B. M., Colombi, P., Jäger, J. & Patrick Lusk, C. Surveillance of nuclear pore complex assembly by ESCRT-III/Vps4. *Cell* **159,** 388–401. doi:10.1016/j.cell.2014.09.012 (2014).

240. Fornasiero, E. F. *et al.* Precisely measured protein lifetimes in the mouse brain reveal differences across tissues and subcellular fractions. *Nature Communications 2018 9:1* **9,** 4230. doi:10.1038/s41467-018-06519-0 (2018).

241. Lasker, K. *et al.* Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. *Proceedings of the National Academy of Sciences of the United States of America* **109,** 1380–1387. doi:10.1073/pnas.1120559109 (2012).

242. Choi, M. *et al.* MSstats: An R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **30,** 2524–2526. doi:10.1093/bioinformatics/btu305 (2014).

243. Schwartz, J. C., Senko, M. W. & Syka, J. E. A two-dimensional quadrupole ion trap mass spectrometer. *Journal of the American Society for Mass Spectrometry* **13,** 659–669. doi:10.1016/S1044-0305(02)00384-7 (2002).

244. De Graaf, E. L. *et al.* Improving SRM assay development: A global comparison between triple quadrupole, Ion trap, and higher energy CID peptide fragmentation spectra. *Journal of Proteome Research* **10,** 4334–4341. doi:10.1021/pr200156b (2011).

245. Shao, C., Zhang, Y. & Sun, W. Statistical characterization of HCD fragmentation patterns of tryptic peptides on an LTQ Orbitrap Velos mass spectrometer. *Journal of Proteomics* **109,** 26–37. doi:10.1016/j.jprot.2014.06.012 (2014).

246. Wysocki, V. H., Tsaprailis, G., Smith, L. L. & Breci, L. A. Mobile and localized protons: A framework for understanding peptide dissociation. *Journal of Mass Spectrometry* **35,** 1399–1406. doi:10.1002/1096-9888(200012)35:12<1399::AID-JMS86>3.0.CO;2-R (2000).

247. Boyd, R. & Somogyi, A. The mobile proton hypothesis in fragmentation of protonated peptides: A perspective. *Journal of the American Society for Mass Spectrometry* **21,** 1275–1278. doi:10.1016/j.jasms.2010.04.017 (2010).

248. Palzs, B. & Suhal, S. Fragmentation pathways of protonated peptides. *Mass Spectrometry Reviews* **24,** 508–548. doi:10.1002/mas.20024 (2005).

249. Pichler, P. *et al.* Peptide labeling with isobaric tags yields higher identification rates using iTRAQ 4-plex compared to TMT 6-plex and iTRAQ 8-plex on LTQ Orbitrap. *Analytical chemistry* **82,** 6549–58. doi:10.1021/ac100890k (2010).

250. Nielsen, M. L., Savitski, M. M. & Zubarev, R. A. Improving Protein Identification Using Complementary Fragmentation Techniques in Fourier Transform Mass Spectrometry. *Molecular & Cellular Proteomics* **4,** 835–845. doi:10.1074/mcp.T400022-MCP200 (2005).

251. Gorshkov, V., Verano-Braga, T. & Kjeldsen, F. SuperQuant: A Data Processing Approach to Increase Quantitative Proteome Coverage. *Analytical Chemistry.* doi:10.1021/acs.analchem.5b01166 (2015).

252. Jiang, X. *et al.* Sensitive and Accurate Quantitation of Phosphopeptides Using TMT Isobaric Labeling Technique. *Journal of Proteome Research* **16,** 4244–4252. doi:10.1021/acs.jproteome.7b00610 (2017).

253. Gabriels, R., Martens, L. & Degroeve, S. Updated MS2PIP web server delivers fast and accurate MS2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *Nucleic Acids Research* **47,** W295–W299. doi:10.1093/nar/gkz299 (2019).

254. Lawson, T. N. *et al.* MsPurity: Automated Evaluation of Precursor Ion Purity for Mass Spectrometry-Based Fragmentation in Metabolomics. *Analytical Chemistry* **89,** 2432–2439. doi:10.1021/acs.analchem.6b04358 (2017).

255. Hogrebe, A. *et al.* Benchmarking common quantification strategies for large-scale phosphoproteomics. *Nature Communications* **9,** 1045. doi:10.1038/s41467-018-03309-6 (2018).

256. Zecha, J. *et al.* Peptide Level Turnover Measurements Enable the Study of Proteoform Dynamics. *Molecular & Cellular Proteomics* **17,** 974–992. doi:10.1074/mcp.RA118.000583 (2018).

257. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476. doi:10.1038/nature07509 (2008).

258. Grinfeld, D. *et al.* Phase-constrained spectrum deconvolution for fourier transform mass spectrometry. *Analytical Chemistry* **89,** 1202–1211. doi:10.1021/acs.analchem.6b03636 (2017).

259. Saba, J. *et al.* Enhanced sensitivity in proteomics experiments using FAIMS coupled with a hybrid linear ion trap/orbitrap mass spectrometer. *Journal of Proteome Research* **8,** 3355–3366. doi:10.1021/pr801106a (2009).

260. Pfammatter, S., Bonneil, E. & Thibault, P. Improvement of Quantitative Measurements in Multiplex Proteomics Using High-Field Asymmetric Waveform Spectrometry. *Journal of Proteome Research* **15,** 4653–4665. doi:10.1021/acs.jproteome.6b00745 (2016).

261. Hebert, A. S. *et al.* Comprehensive Single-Shot Proteomics with FAIMS on a Hybrid Orbitrap Mass Spectrometer. *Analytical Chemistry* **90,** 9529–9537. doi:10.1021/acs.analchem.8b02233 (2018).

262. Bekker-Jensen, D. B. *et al.* A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients. *bioRxiv,* 860643. doi:10.1101/860643 (2019).

263. Nahnsen, S., Bielow, C., Reinert, K. & Kohlbacher, O. Tools for label-free peptide quantification. *Molecular and Cellular Proteomics* **12,** 549–556. doi:10.1074/mcp.R112.025163 (2013).

264. Geib, T. *et al.* Triple Quadrupole Versus High Resolution Quadrupole-Time-of-Flight Mass Spectrometry for Quantitative LC-MS/MS Analysis of 25-Hydroxyvitamin D in Human Serum. *Journal of the American Society for Mass Spectrometry* **27,** 1404–1410. doi:10.1007/s13361-016-1412-2 (2016).

265. Ting, Y. S. *et al.* PECAN: Library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nature Methods* **14,** 903–908. doi:10.1038/nmeth.4390 (2017).

266. Tsou, C. C. *et al.* DIA-Umpire: Comprehensive computational framework for data-independent acquisition proteomics. *Nature Methods* **12,** 258–264. doi:10.1038/nmeth.3255 (2015).

267. Bruderer, R., Bernhardt, O. M., Gandhi, T. & Reiter, L. High-precision iRT prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation. *Proteomics* **16,** 2246–2256. doi:10.1002/pmic.201500488 (2016).

268. Brenes, A., Hukelmann, J., Bensaddek, D. & Lamond, A. I. Multibatch TMT Reveals False Positives, Batch Effects and Missing Values. *Molecular & cellular proteomics : MCP* **18,** 1967–1980. doi:10.1074/mcp.RA119.001472 (2019).

Part VI

Acknowledgment

I would first of all like to thank Mikhail Savitski whose enthusiasm and dedication to the development of proteomic software spurred my decision to start on the road towards undertaking this project. The invaluable discussions and time spent mentoring and slowly encouraging my progress cannot be overestimated. I am also thankful to Bernhard Küster for agreeing to oversee this entire pursuit - the right discussions at the right moments have steered me in the right direction.

I am very much indebted to Marcus Bantscheff, Gitte Neubauer and the Cellzome leadership team for allowing me the opportunity to pursue a PhD project alongside my daily work. It must have been clear that this would by no means be an easy journey considering the challenge of managing a full time job whilst maintaining focus on a different, yet related, project. I would like to extend special thanks to my close departmental colleagues at Cellzome, a GSK company, without whose help I could not have progressed so far. There from the very beginning was Gavain Sweetman and Holger Franken and, in the last year or two, Christian Fufezan. To you I am truly obliged for the insightful, interesting and informative discussions to shape my mind and improve my code.

It is hard to express the amount of thanks and appreciation that I have to express to Laura who was with me for so much of this odyssey, and who was indirectly affected by many of its ups and downs. Without her continual love, support and determined encouragement, I honestly doubt I would have made it to this point.

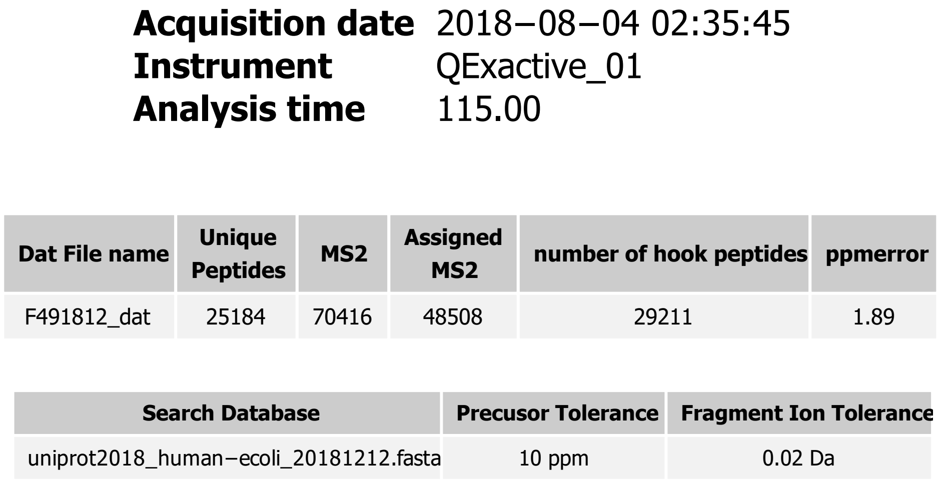
Finally, I thank my parents and family. Their quiet inspiration and backing has accompanied me throughout and has ensured that I made it to the finish line.

Part VII

Appendix

A Description of R-script output

Below is a description of the set of simple QC outputs from a single instrument run generated using an R-script which is part of the isobarQuant suite. The data is entirely found in the .hdf5 file name after the .raw file which was processed. The individual plots are also described in table 4, on page 76

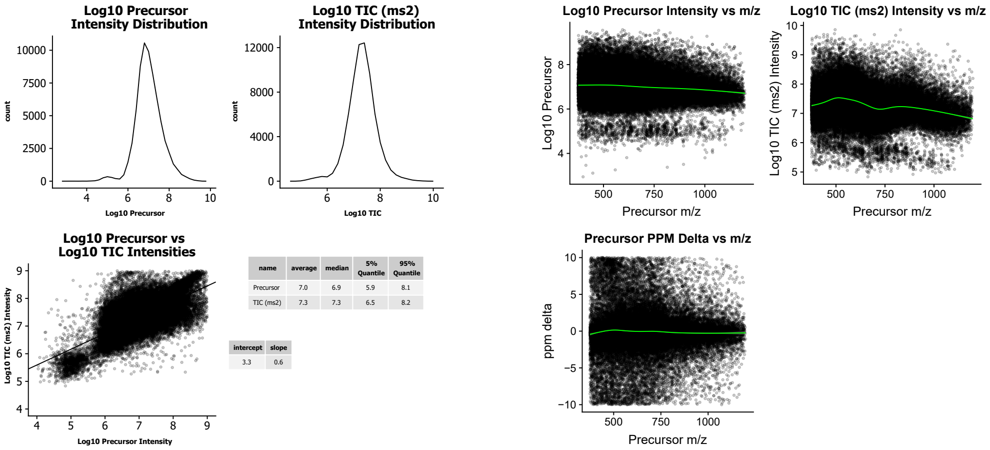


Acquisition date	2018−08−04 02:35:45
Instrument	QExactive_01
Analysis time	115.00

Dat File name	Unique Peptides	MS2	Assigned MS2	number of hook peptides	ppmerror
F491812_dat	25184	70416	48508	29211	1.89

Search Database	Precusor Tolerance	Fragment Ion Tolerance
uniprot2018_human−ecoli_20181212.fasta	10 ppm	0.02 Da

Figure (1): Example of QC overview page created via an R-script for an example run. A small excerpt of the total parameters and metrics stored in the .hdf5 file are shown to give the user a basic idea of the performance of the run.



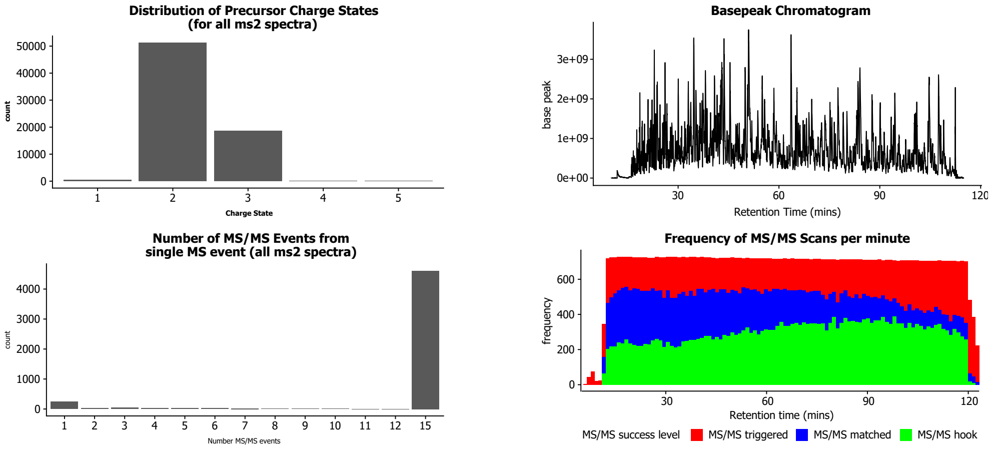
name	average	median	5% Quantile	95% Quantile
Precursor	7.0	6.9	5.9	8.1
TIC (ms2)	7.3	7.3	6.5	8.2

intercept	slope
3.3	0.6

(a) Distributions of precursor intensities and TICs for MS1 and MS2 spectra respectively.

(b) Top panel: Log_{10}-transformed precursor intensity and TIC is plotted for all MS1 and MS2 spectra. Lower panel: deviation of precursor *m/z* from expected value.

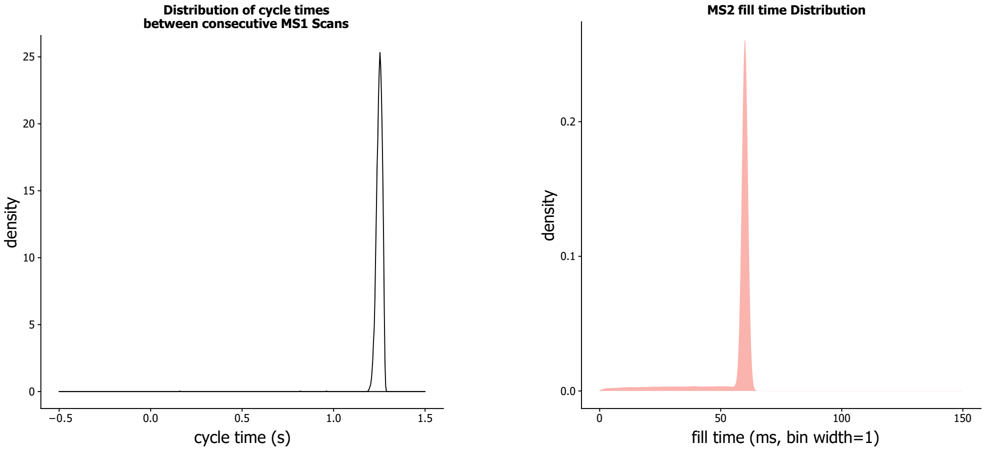
Figure (2): QC plots relating to MS1 and MS2 intensities can help the experimenter determine problems in signal transmission within and between mass spec runs. The parts-per-million (ppm) difference between measured and theoretical precursor *m/z* can yield insight into how well calibrated the instrument is.



(a) Top panel: Histogram of precursor charge states throughout run. Lower panel: Number of MS/MS events triggered for a single precursor

(b) Top panel: Basepeak chromatogram of entire MS run in single minute bins. Lower panel: Frequency of success rate of MS/MS spectra for the whole MS run divided into single minute RT bins. Red signifies no match found by the search engine. Blue bars highlight the frequency of matches of the MS/MS spectrum to a peptide and green, high-quality PSM matches.

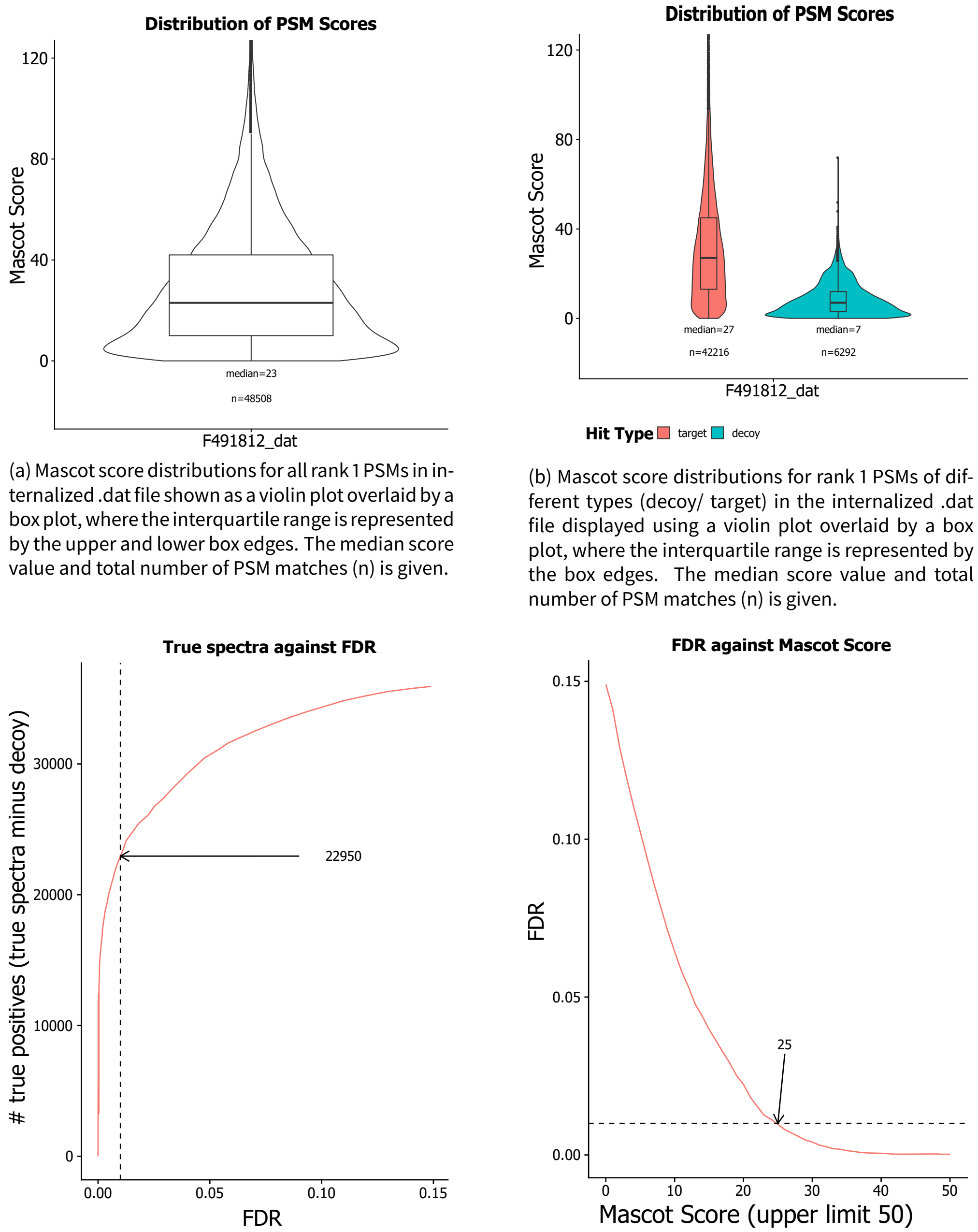
Figure (3): QC plots relating to charge state, frequency and success of triggered MS/MS events per precursor and a profile of the chromatographic baseline peaks across the run.



(a) Density plot distribution of cycle times between consecutive MS1 scans

(b) Density plot of Orbitrap fill times for all MS/MS fragments

Figure (4): QC plots relating to the time between subsequent MS Scans and time taken to fill the Orbitrap with MS/MS fragments. Differences between run or compared to a known standard can be a useful metric for experimenters to look into instrument performance.



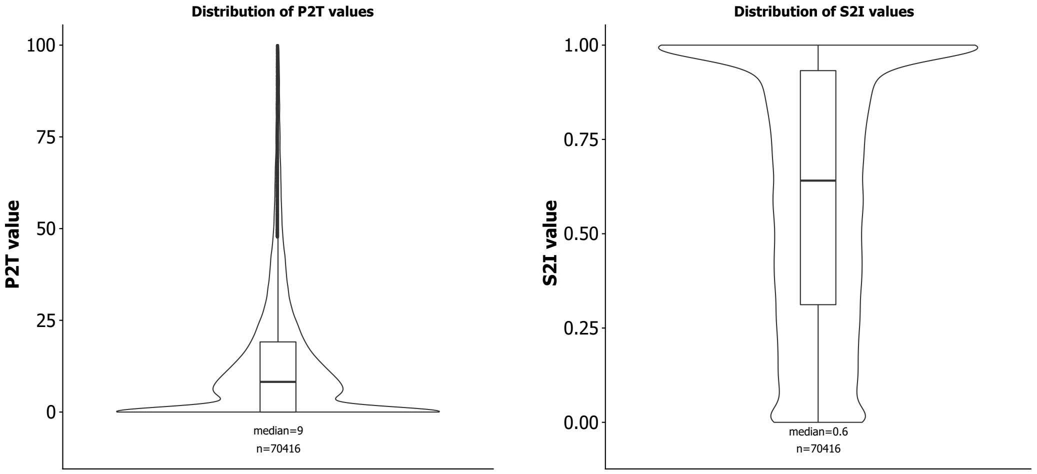
(a) Mascot score distributions for all rank 1 PSMs in internalized .dat file shown as a violin plot overlaid by a box plot, where the interquartile range is represented by the upper and lower box edges. The median score value and total number of PSM matches (n) is given.

(b) Mascot score distributions for rank 1 PSMs of different types (decoy/ target) in the internalized .dat file displayed using a violin plot overlaid by a box plot, where the interquartile range is represented by the box edges. The median score value and total number of PSM matches (n) is given.

(c) ROC curve showing numbers of cumulative true positive PSMs against estimated FDR. Number of 'true spectra' (non-decoy PSMs) passing 1% FDR threshold is also displayed

(d) Plot of cumulative FDR for Mascot scores of PSMs to yield the corresponding q-value. The q-value equaling an FDR of 1% is marked.

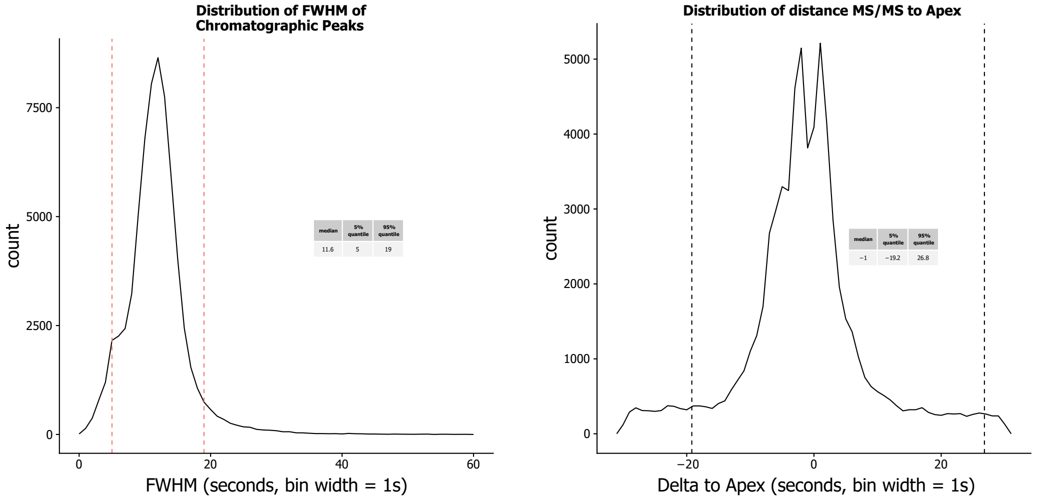
Figure (5): QC plots relating to Mascot scoring and FDRof the processed .raw data give immediate feedback to experimenters about the performance of the instrument in terms of identified PSMs



(a) Violin overlaid withe box plot (with IQR corresponding to upper and lower box-edges) to display P2T distribution for all precursor spectra. Median and total MS/MS counts are also displayed.

(b) Violin overlaid with box plot (with IQR corresponding to upper and lower box-edges) to display S2I distribution for all precursor spectra. Median and total MS/MS counts are also displayed

Figure (6): QC plots relating to distribution of calculated P2T noise level and precursor interference S2I for ions selected for fragmentation to MS/MS spectra



(a) Distribution of FWHM values for chromatographic peaks within MS run. Upper and lower quantile values (5% / 95%) also given.

(b) Frequency distribution of time difference between apex of XIC and triggered MS/MS event

Figure (7): QC plots relating to LC performance and peak picking compared to chromatographic apex of precursors' XIC for the given run

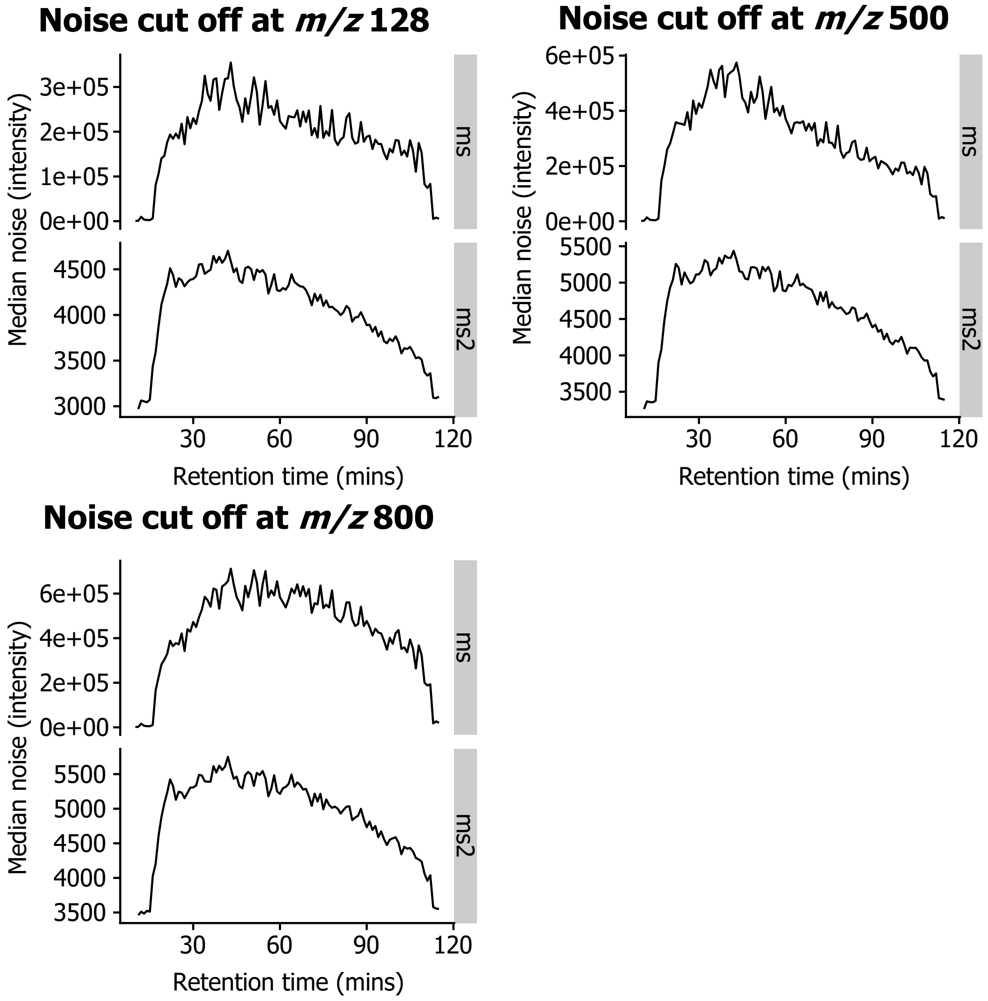


Figure (8): Distributions of mean noise intensity measured for MS1 and MS2 scans at three different *m/z* values at time points throughout the entire run.

B Location of code

It would be impractical to append all code within the isobarQuant suite to the body of this thesis. The code is freely available for download or inspection at the Github repository see https://github.com/protcode/isob or for the zipped downloadable bundle https://github.com/protcode/isob/archive/master.zip